Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 59 (2019)

Samplers and Extractors for Unbounded Functions

Rohit Agrawal*

September 17, 2019

Abstract

Błasiok (SODA'18) recently introduced the notion of a subgaussian sampler, defined as an averaging sampler for approximating the mean of functions $f : \{0,1\}^m \to \mathbb{R}$ such that $f(U_m)$ has subgaussian tails, and asked for explicit constructions. In this work, we give the first explicit constructions of subgaussian samplers (and in fact averaging samplers for the broader class of subexponential functions) that match the best known constructions of averaging samplers for $[0,1]$-bounded functions in the regime of parameters where the approximation error ε and failure probability δ are subconstant. Our constructions are established via an extension of the standard notion of randomness extractor (Nisan and Zuckerman, JCSS'96) where the error is measured by an arbitrary divergence rather than total variation distance, and a generalization of Zuckerman's equivalence (Random Struct. Alg.'97) between extractors and samplers. We believe that the framework we develop, and specifically the notion of an extractor for the Kullback–Leibler (KL) divergence, are of independent interest. In particular, KL-extractors are stronger than both standard extractors and subgaussian samplers, but we show that they exist with essentially the same parameters (constructively and non-constructively) as standard extractors.

1 Introduction

1.1 Averaging samplers

Averaging (or oblivious) samplers, introduced by Bellare and Rompel [BR94], are one of the main objects of study in pseudorandomness. Used to approximate the mean of a $[0,1]$-valued function with minimal randomness and queries, an averaging sampler takes a short random seed and produces a small set of correlated points such that any given $[0,1]$-valued function will (with high probability) take approximately the same mean on these points as on the entire space. Formally,

Definition 1.1 ([BR94])**.** A function $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ is a (δ,ε) *averaging sampler* if for all $f : \{0,1\}^m \to [0,1]$, it holds that

$$\Pr_{x\sim U_n}\left[\left|\frac{1}{D}\sum_{i=1}^{D} f(\mathrm{Samp}(x)_i) - \mathbb{E}[f(U_m)]\right| > \varepsilon\right] \le \delta,$$

where U_n is the uniform distribution on $\{0,1\}^n$. The number n is the *randomness complexity* of the sampler, and D is the *sample complexity*. A sampler is *explicit* if $\mathrm{Samp}(x)_i$ can be computed in time $\mathrm{poly}(n, m, \log D)$.

Traditionally, averaging samplers have been used in the context of randomness-efficient error reduction for algorithms and protocols, where the function f is the indicator of a set ($\{0,1\}$-valued), or more generally the acceptance probability of an algorithm or protocol ($[0,1]$-valued). There has been significant effort in the literature to establish optimal explicit and non-explicit constructions of samplers, which we summarize in

*John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA. Email: rohitagr@seas.harvard.edu. Supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

ISSN 1433-8092

Table 1. We recommend the survey of Goldreich [Gol11b] for more details, especially regarding non-averaging samplers[1].

Table 1: Best known constructions of averaging samplers for $[0,1]$-valued functions

Key Idea	Randomness complexity n	Sample complexity D	Best regime
Pairwise-independent Expander Neighbors [GW97]	$m + O(\log(1/\delta) + \log(1/\varepsilon))$	$O\left(\frac{1}{\delta\varepsilon^2}\right)$	$\delta = \Omega(1)$
Ramanujan Expander Neighbors[a] [KPS85, GW97]	m	$O\left(\frac{1}{\delta\varepsilon^2}\right)$	$\delta = \Omega(1)$
Extractors [Zuc97, GW97, RVW00, GUV09]	$m + (1+\alpha) \cdot \log(1/\delta)$ any constant $\alpha > 0$	$\mathrm{poly}(\log(1/\delta), 1/\varepsilon)$	$\varepsilon, \delta = o(1)$
Expander Walk Chernoff [Gil98]	$m + O(\log(1/\delta)/\varepsilon^2)$	$O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$	$\varepsilon = \Omega(1)$
Pairwise Independence [CG89]	$O(m)$	$O\left(\frac{1}{\delta\varepsilon^2}\right)$	None, but simple
Non-Explicit [Zuc97]	$m + \log(1/\delta) - \log\log(1/\delta) + O(1)$	$O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$	All
Lower Bound [CEG95, Zuc97, RT00]	$m + \log(1/\delta) + \log(1/\varepsilon) - \log(D) - O(1)$	$\Omega\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$	N/A

[a] Requires explicit constructions of Ramanujan graphs.

However, averaging samplers can also have uses beyond bounded functions: Błasiok [Bła18b], motivated by an application in streaming algorithms, introduced the notion of a *subgaussian sampler*, which he defined as an averaging sampler for functions $f : \{0,1\}^m \to \mathbb{R}$ such that $f(U_m)$ is a subgaussian random variable. Since subgaussian random variables have strong tail bounds, subgaussian functions from $\{0,1\}^m$ have a range contained in an interval of length $O(\sqrt{m})$, and thus one can construct a subgaussian sampler from a $[0,1]$-sampler by simply scaling the error ε by a factor of $O(\sqrt{m})$. Unfortunately, looking at Table 1 one sees that this induces a multiplicative dependence on m in the sample complexity, and for the expander walk sampler induces a dependence of $m\log(1/\delta)$ in the randomness complexity. This loss can be avoided for some samplers, such as the sampler of Chor and Goldreich [CG89] based on pairwise independence (as its analysis requires only bounded variance) and (as we will show) the Ramanujan Expander Neighbor sampler of [KPS85, GW97], but Błasiok showed [Bła18a] that the expander-walk sampler does not in general act as a subgaussian sampler without reducing the error to $o(1)$. We remark briefly that the median-of-averages sampler of Bellare, Goldreich, and Goldwasser [BGG93] still works and is optimal up to constant factors in the subgaussian setting (since the underlying pairwise independent sampler works), but it is not an averaging sampler[1], and matching its parameters with an averaging sampler remains open in general even for $[0,1]$-valued functions.

One of the contributions of this work is to give explicit averaging samplers for subgaussian functions (in fact even for *subexponential* functions that satisfy weaker tail bounds) matching the extractor-based samplers for $[0,1]$-valued functions in Table 1 (up to the hidden polynomial in the sample complexity). This achieves the best parameters currently known in the regime of parameters where ε and δ are both subconstant, and in particular has no dependence on m in the sample complexity. We also show non-constructively that subexponentially samplers exist with essentially the same parameters as $[0,1]$-valued samplers.

Theorem 1.2 (Informal version of Theorem 6.1 and Corollary 6.7)**.** *For every integer $m \in \mathbb{N}$, $1 > \delta, \varepsilon > 0$, and $\alpha > 0$, there is a function* $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ *that is:*

- *an explicit subgaussian (in fact subexponential) sampler with randomness complexity $n = m + (1+\alpha) \cdot \log(1/\delta)$ and sample complexity $D = \mathrm{poly}(\log(1/\delta), 1/\varepsilon)$ (see Theorem 6.1)*

[1] A non-averaging sampler is an algorithm Samp which makes oracle queries to f and outputs an estimate of its average which is good with high probability, but need not simply output the average of f's values on the queried points.

- *a non-constructive subexponential sampler with randomness complexity* $n = m + \log(1/\delta) - \log\log(1/\delta) + O(1)$ *and sample complexity* $D = O(\log(1/\delta)/\varepsilon^2)$ *(see Corollary 6.7).*

1.2 Randomness extractors

To prove Theorem 1.2, we develop a corresponding theory of generalized *randomness extractors* which we believe is of independent interest. For bounded functions, Zuckerman [Zuc97] showed that averaging samplers are essentially equivalent to randomness extractors, and in fact several of the best-known constructions of such samplers arose as extractor constructions. Formally, a randomness extractor is defined as follows:

Definition 1.3 (Nisan and Zuckerman [NZ96])**.** A function $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is said to be a (k, ε) *extractor* if for every distribution X over $\{0,1\}^m$ satisfying $\max_{x\in\{0,1\}^n} \Pr[X = x] \leq 2^{-k}$, the distributions $\mathrm{Ext}(X, U_d)$ and U_m are ε-close in total variation distance. Equivalently, for all $f : \{0,1\}^m \to [0,1]$ it holds that $\mathbb{E}[f(\mathrm{Ext}(X, U_d))] - \mathbb{E}[f(U_m)] \leq \varepsilon$. The number d is called the *seed length*, and m the *output length*.

The formulation of Definition 1.3 in terms of $[0,1]$-valued functions implies that extractors produce an output distribution that is indistinguishable from uniform by all bounded functions f. It is therefore natural to consider a variant of this definition for a different set $\mathcal{F}$ of test functions $f : \{0,1\}^m \to \mathbb{R}$ which need not be bounded.

Definition 1.4 (Special case of Definition 3.1 using Definition 2.5)**.** A function $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is said to be a (k, ε) extractor for a set of real-valued functions $\mathcal{F}$ from $\{0,1\}^m$ if for every distribution X over $\{0,1\}^m$ satisfying $\max_{x\in\{0,1\}^n} \Pr[X = x] \leq 2^{-k}$ and every $f \in \mathcal{F}$, it holds that $\mathbb{E}[f(\mathrm{Ext}(X, U_d))] - \mathbb{E}[f(U_m)] \leq \varepsilon$.

We show that much of the theory of extractors and samplers carries over to this more general setting. In particular, we generalize the connection of Zuckerman [Zuc97] to show that extractors for a class of functions of $\mathcal{F}$ are also samplers for that class, along with the converse (though as for total variation distance, there is some loss of parameters in this direction). Thus, to construct a subgaussian sampler it suffices (and is preferable) to construct a corresponding extractor for subgaussian test functions, which is how we prove Theorem 1.2.

Unfortunately, the distance induced by subgaussian test functions is not particularly pleasant to work with: for example the point masses on 0 and 1 in $\{0,1\}$ are $O(1)$ apart, but embedding them in the larger universe $\{0,1\}^m$ leads to distributions which are $\Theta(\sqrt{m})$ apart. We solve this problem by constructing extractors for a stronger notion, the *Kullback–Leibler (KL) divergence*, equivalently, extractors whose output is required to have very high Shannon entropy.

Definition 1.5 (Special case of Definition 3.1 using KL divergence)**.** A function $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is said to be a (k, ε) KL-extractor if for every distribution X over $\{0,1\}^m$ satisfying $\max_{x\in\{0,1\}^n} \Pr[X = x] \leq 2^{-k}$ it holds that $\mathrm{KL}(\mathrm{Ext}(X, U_d) \parallel U_m) \leq \varepsilon$, or equivalently $\mathrm{H}(\mathrm{Ext}(X, U_d)) \geq m - \varepsilon$.

A strong form of Pinsker's inequality (e.g. [BLM13, Lemma 4.18]) implies that a (k, ε^2) KL-extractor is also a (k, ε) extractor for subgaussian test functions. The KL divergence has the advantage that is nonincreasing under the application of functions (the famous *data-processing inequality*), and although it does not satisfy a traditional triangle inequality, it does satisfy a similar inequality when one of the segments satisfies stronger ℓ_2 bounds. These properties allow us to show that the zig-zag product for extractors of Reingold, Wigderson, and Vadhan [RVW00] also works for KL-extractors, and therefore to construct KL-extractors with seed length depending on n and k only through the *entropy deficiency* $n - k$ of X rather than n itself, which in the sampler perspective corresponds to a sampler with sample complexity depending on the failure probability δ rather than the universe size 2^m. Hence, we prove Theorem 1.2 by constructing corresponding KL-extractors.

Theorem 1.6 (Informal version of Theorem 6.2)**.** *For all integers* m*,* $1 > \delta, \varepsilon > 0$*, and* $\alpha > 0$ *there is an explicit* (k, ε) *KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with* $n = m + (1 + \alpha) \cdot \log(1/\delta)$*,* $k = n - \log(1/\delta)$*, and* $d = O(\log\log(1/\delta) + \log(1/\varepsilon))$*.*

Though the above theorem is most interesting in the high min-entropy regime where $n - k = o(n)$, we also show the existence of KL-extractors matching most of the existing constructions of total variation extractors. In particular, we note that extractors for ℓ_2 are immediately KL-extractors without loss of parameters, and also that any extractor can be made a KL-extractor by taking slightly smaller error, so that the extractors of Guruswami, Umans, and Vadhan [GUV09] can be taken to be KL-extractors with essentially the same parameters.

Furthermore, in addition to our explicit constructions, we also show non-constructively that KL-extractors (and hence subgaussian extractors) exist with very good parameters:

Theorem 1.7 (Informal version of Theorem 5.30)**.** *For any integers $k < n \in \mathbb{N}$ and $1 > \varepsilon > 0$ there is a (k, ε) KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with $d = \log(n-k) + \log(1/\varepsilon) + O(1)$ and $m = k + d - \log(1/\varepsilon) - O(1)$.*

One key thing to note about the nonconstructive KL extractors of the above theorem is that they incur an entropy loss of only $1 \cdot \log(1/\varepsilon)$, whereas total variation extractors necessarily incur entropy loss $2 \cdot \log(1/\varepsilon)$ by the lower bound of Radhakrishnan and Ta-Shma [RT00]. In particular, by Pinsker's inequality, (k, ε^2) KL-extractors with the above parameters are also optimal (k, ε) standard (total variation) extractors [RT00], so that one does not lose anything by constructing a KL-extractor rather than a total variation extractor. We also remark that the above theorem gives subgaussian samplers with better parameters than a naive argument that a random function should directly be a subgaussian sampler, as it avoids the need to take a union bound over $O(M^M) = O(2^{M \log M})$ test functions (for $M = 2^m$) which results in additional additive $\log\log$ factors in the randomness complexity.

In the total variation setting, there are only a couple of methods known to explicitly achieve optimal entropy loss $2 \cdot \log(1/\varepsilon)$, the easiest of which is to use an extractor which natively has this sort of loss, of which only three are known: An extractor from random walks over Ramanujan Graphs due to Goldreich and Wigderson [GW97], the Leftover Hash Lemma due to Impagliazzo, Levin, and Luby [ILL89] (see also [McI87, BBR88]), and the extractor based on almost-universal hashing of Srinivasan and Zuckerman [SZ99]. Unfortunately, all of these are ℓ_2 extractors and so must have seed length linear in $\min(n-k, m)$ (cf. [Vad12, Problem 6.4]), rather than logarithmic in $n - k$ as known non-constructively. The other alternative is to use the generic reduction of Raz, Reingold, and Vadhan [RRV02] which turns any extractor Ext with entropy loss Δ into one with entropy loss $2 \cdot \log(1/\varepsilon) + O(1)$ by paying an additive $O(\Delta + \log(n/\varepsilon))$ in seed length. We show that all of these ℓ_2 extractors and the [RRV02] transformation also work to give KL-extractors with entropy loss $1 \cdot \log(1/\varepsilon) + O(1)$, so that applications which require minimal entropy loss can also use explicit constructions of KL-extractors.

1.3 Future directions

Broadly speaking, we hope that the perspective of KL-extractors will bring new tools (perhaps from information theory) to the construction of extractors and samplers. For example, since KL-extractors can have seed length with dependence on ε of only $1 \cdot \log(1/\varepsilon)$, trying to explicitly construct a KL-extractor with seed length $1 \cdot \log(1/\varepsilon) + o(\min(n-k, k))$ may also shed light on how to achieve optimal dependence on ε in the total variation setting.

In the regime of constant $\varepsilon = \Omega(1)$, we do not have explicit constructions of subgaussian samplers matching the expander-walk sampler of Gillman [Gil98] for $[0,1]$-valued functions, which achieves randomness complexity $m + O(\log(1/\delta))$ and sample complexity $O(\log(1/\delta))$, as asked for by Błasiok [Bła18b]. From the extractor point-of-view, it would suffice (by the reduction of [GW97, RVW00] that we analyze for KL-extractors) to construct explicit *linear degree* KL-extractors with parameters matching the linear degree extractor of Zuckerman [Zuc07], i.e. with seed length $d = \log(n) + O(1)$ and $m = \Omega(k)$ for $\varepsilon = \Omega(1)$. A potentially easier problem, since the Zuckerman linear degree extractor is itself based on the expander-walk sampler, could be to instead match the parameters of the near-linear degree extractors of Ta-Shma, Zuckerman, and Safra [TZS06] based on Reed–Muller codes, thereby achieving sample complexity $O(\log(1/\delta) \cdot \mathrm{poly}\log\log(1/\delta))$.

Finally, we hope that KL-extractors can also find uses beyond being subgaussian samplers and total variation extractors: for example it seems likely that there are applications (perhaps in coding or cryptography, cf. [BDK+11]) where it is more important to have high Shannon entropy in the output than small total variation distance to uniform, in which case one may be able to use (k, ε) KL-extractors with entropy loss only $1 \cdot \log(1/\varepsilon)$ directly, rather than a total variation extractor or (k, ε^2) KL-extractor with entropy loss $2 \cdot \log(1/\varepsilon)$.

2 Preliminaries

2.1 (Weak) statistical divergences and metrics

Our results in general will require very few assumptions on notions of "distance" between probability distributions, so we will give a general definition and indicate in our theorems when we need which assumptions.

Definition 2.1. A *weak statistical divergence* (or simply *weak divergence*) on a finite set $\mathcal{X}$ is a function D from pairs of probability distributions over X to $\mathbb{R} \cup \{\pm\infty\}$. We write $\mathrm{D}(P \parallel Q)$ for the value of D on distributions P and Q. Furthermore

1. If $\mathrm{D}(P \parallel Q) \geq 0$ with equality iff $P = Q$, then D is *positive-definite*, and we simply call D a *divergence*.
2. If $\mathrm{D}(P \parallel Q) = \mathrm{D}(Q \parallel P)$, then D is *symmetric*.
3. If $\mathrm{D}(P \parallel R) \leq \mathrm{D}(P \parallel Q) + \mathrm{D}(Q \parallel R)$, then D satisfies the *triangle inequality*.
4. If $\mathrm{D}(\lambda P_1 + (1-\lambda) P_2 \parallel \lambda Q_1 + (1-\lambda) Q_2) \leq \lambda \, \mathrm{D}(P_1 \parallel Q_1) + (1-\lambda) \, \mathrm{D}(P_2 \parallel Q_2)$ for all $\lambda \in [0,1]$, then D is *jointly convex*. If this holds only when $Q_1 = Q_2$ then D is *convex in its first argument*.
5. If D is defined on all finite sets $\mathcal{Y}$ and for all functions $f : \mathcal{X} \to \mathcal{Y}$ the divergence is nonincreasing under f, that is $\mathrm{D}(f(P) \parallel f(Q)) \leq \mathrm{D}(P \parallel Q)$, then D satisfies the *data-processing inequality*.

If D is positive-definite, symmetric, and satisfies the triangle inequality, then it is called a *metric*.

Example 2.2. The ℓ_p *distance* for $p > 0$ between probability distributions over $\mathcal{X}$ is

$$d_{\ell_p}(P, Q) \stackrel{\text{def}}{=} \left(\sum_{x \in \mathcal{X}} \left| P_x - Q_x \right|^p \right)^{1/p}$$

and is positive-definite and symmetric. Furthermore, for $p \geq 1$ it satisfies the triangle inequality (and so is a metric), and is jointly convex. The ℓ_p distance is nonincreasing in p.

Example 2.3. The *total variation distance* is

$$d_{TV}(P, Q) \stackrel{\text{def}}{=} \frac{1}{2} d_{\ell_1}(P, Q) = \sup_{S \subseteq \mathcal{X}} \left| \Pr[P \in S] - \Pr[Q \in S] \right| = \sup_{f \in [0,1]^{\mathcal{X}}} \Big(\mathbb{E}[f(P)] - \mathbb{E}[f(Q)] \Big)$$

and is a jointly convex metric that satisfies the data-processing inequality.

Example 2.4 (Rényi Divergences [Rén61])**.** For two probability distributions P and Q over a finite set $\mathcal{X}$, the *Rényi α-divergence* or *Rényi divergence of order α* is defined for real $0 < \alpha \neq 1$ by

$$\mathrm{D}_\alpha(P \parallel Q) \stackrel{\text{def}}{=} \frac{1}{\alpha - 1} \log \left(\sum_{x \in \mathcal{X}} \frac{P_x^\alpha}{Q_x^{\alpha - 1}} \right)$$

where the logarithm is in base 2 (as are all logarithms in this paper unless noted otherwise). The Rényi divergence is continuous in α and so is defined by taking limits for $\alpha \in \{0, 1, \infty\}$, giving for $\alpha = 0$ the divergence $\mathrm{D}_0(P \parallel Q) \stackrel{\text{def}}{=} \log(1/\Pr_{x \sim Q}[P_x \neq 0])$, for $\alpha = 1$ the *Kullback–Leibler (or KL) divergence*

$$\mathrm{KL}(P \parallel Q) \stackrel{\text{def}}{=} \mathrm{D}_1(P \parallel Q) = \sum_{x \in X} P_x \log \frac{P_x}{Q_x},$$

and for $\alpha = \infty$ the *max-divergence* $\mathrm{D}_\infty(P \parallel Q) \stackrel{\text{def}}{=} \max_{x \in X} \log \frac{P_x}{Q_x}$. The Rényi divergence is nondecreasing in α. Furthermore, when $\alpha \le 1$ the Rényi divergence is jointly convex, and for all α the Rényi divergence satisfies the data-processing inequality [vEH14].

When $Q = U_\mathcal{X}$ is the uniform distribution over the set $\mathcal{X}$, then for all α, $\mathrm{D}_\alpha(P \parallel U_\mathcal{X}) = \log|\mathcal{X}| - \mathrm{H}_\alpha(P)$ where $0 \le \mathrm{H}_\alpha(P) \le \log|\mathcal{X}|$ is called the *Rényi α-entropy of P*. For $\alpha = 0$, $\mathrm{H}_0(P) = \log|\mathrm{Supp}(P)|$ is the *max-entropy of P*, for $\alpha = 1$, $\mathrm{H}_1(P) = \sum_{x \in \mathcal{X}} P_x \log(1/P_x)$ is the *Shannon entropy of P*, and for $\alpha = \infty$, $\mathrm{H}_\infty(P) = \min_{x \in \mathcal{X}} \log(1/P_x)$ is the *min-entropy of P*.

For $\alpha = 2$, the Rényi 2-entropy can be expressed in terms of the ℓ_2-distance to uniform:

$$\log|\mathcal{X}| - H_2(P) = \mathrm{D}_2(P \parallel U_\mathcal{X}) = \log\bigl(1 + |\mathcal{X}| \cdot d_{\ell_2}(P, U_\mathcal{X})^2\bigr)$$

2.2 Statistical weak divergences from test functions

Zuckerman's connection [Zuc97] between samplers for bounded functions and extractors for total variation distance is based on the following standard characterization of total variation distance as the maximum distinguishing advantage achieved by bounded functions,

$$d_{TV}(P, Q) = \sup_{f \in [0,1]^\mathcal{X}} \mathbb{E}[f(P)] - \mathbb{E}[f(Q)].$$

By considering an arbitrary class of functions in the supremum, we get the following weak divergence:

Definition 2.5. Given a finite $\mathcal{X}$ and a set of real-valued functions $\mathcal{F} \subseteq \mathbb{R}^\mathcal{X}$, the *$\mathcal{F}$-distance on $\mathcal{X}$* between probability measures on $\mathcal{X}$ is denoted by $\mathrm{D}^\mathcal{F}$ and is defined as

$$\mathrm{D}^\mathcal{F}(P \parallel Q) \stackrel{\text{def}}{=} \sup_{f \in \mathcal{F}} \Bigl(\mathbb{E}[f(P)] - \mathbb{E}[f(Q)]\Bigr) = \sup_{f \in \mathcal{F}} \mathrm{D}^{\{f\}}(P \parallel Q),$$

where we use a superscript to avoid confusion with the Csiszár-Morimoto-Ali-Silvey f-divergences [Csi63, Mor63, AS66].

We call the set of functions $\mathcal{F}$ *symmetric* if for all $f \in \mathcal{F}$ there is $c \in \mathbb{R}$ and $g \in \mathcal{F}$ such that $g = c - f$, and *distinguishing* if for all $P \ne Q$ there exists $f \in \mathcal{F}$ with $\mathrm{D}^{\{f\}}(P \parallel Q) > 0$.

Example 2.6. If $\mathcal{F} = \{0,1\}^\mathcal{X}$ or $\mathcal{F} = [0,1]^\mathcal{X}$, then $\mathrm{D}^\mathcal{F}$ is exactly the total variation distance.

Remark 2.7. An equivalent definition of $\mathcal{F}$ being symmetric is that for all $f \in \mathcal{F}$ there exists $g \in \mathcal{F}$ with $\mathrm{D}^{\{g\}}(P \parallel Q) = -D^{\{f\}}(P \parallel Q) = \mathrm{D}^{\{f\}}(Q \parallel P)$ for all distributions P and Q. Hence, one might also consider a weaker notion of symmetry that reverses quantifiers, where $\mathcal{F}$ is "weakly-symmetric" if for all $f \in \mathcal{F}$ and distributions P and Q there exists $g \in \mathcal{F}$ such that $\mathrm{D}^{\{g\}}(P \parallel Q) = -\mathrm{D}^{\{f\}}(P \parallel Q) = \mathrm{D}^{\{f\}}(Q \parallel P)$. However, such a class $\mathcal{F}$ gives exactly the same weak divergence $\mathrm{D}^\mathcal{F}$ as its "symmetrization" $\overline{\mathcal{F}} = \mathcal{F} \cup \{-f \mid f \in \mathcal{F}\}$, so we do not need to introduce this more complex notion.

Remark 2.8. By identifying distributions with their probability mass function, one can realize $\mathbb{E}[f(P)] - \mathbb{E}[f(Q)]$ as an inner product $\langle P - Q, f\rangle$. Definition 2.5 can thus be written as $\mathrm{D}^\mathcal{F}(P \parallel Q) = \sup_{f \in \mathcal{F}} \langle P - Q, f\rangle$, which is essentially the notion of indistinguishability considered in several prior works, (see e.g. the survey of Reingold, Trevisan, Tulsiani, and Vadhan [RTTV08]), but without requiring all f to be bounded.

Remark 2.9. For simplicity, all our probabilistic distributions are given only for random variables and distributions over finite sets as this is all we need for our application. A more general version of Definition 2.5 has been studied by e.g. Zolotarev [Zol84] and Müller [Mül97] and is commonly used in developments of Stein's method in probability.

We now establish some basic properties of $\mathrm{D}^\mathcal{F}$.

Lemma 2.10. *Let $\mathcal{F} \subseteq \mathbb{R}^\mathcal{X}$ be a set of real-valued functions over a finite set $\mathcal{X}$. Then $\mathrm{D}^\mathcal{F}$ satisfies the triangle inequality and is jointly convex, and*

1. *if $\mathcal{F}$ is symmetric then $\mathrm{D}^{\mathcal{F}}$ is symmetric and*

$$\mathrm{D}^{\mathcal{F}}(P \parallel Q) = \sup_{f \in \mathcal{F}} \Big| \mathbb{E}[f(P)] - \mathbb{E}[f(Q)] \Big| \geq 0,$$

2. *if $\mathcal{F}$ is distinguishing then $\mathrm{D}^{\mathcal{F}}$ is positive-definite,*

so that if $\mathcal{F}$ is both symmetric and distinguishing then $\mathrm{D}^{\mathcal{F}}$ is a jointly convex metric on probability distributions over $\mathcal{X}$, in which case we also use the notation $d_{\mathcal{F}}(P,Q) \stackrel{\text{def}}{=} \mathrm{D}^{\mathcal{F}}(P \parallel Q)$.

Proof. The triangle inequality and joint convexity both follow from the linearity of each $\mathrm{D}^{\{f\}}$, as by linearity of expectation, for all $f : \mathcal{X} \to \mathbb{R}$ it holds that

$$\mathrm{D}^{\{f\}}(P \parallel R) = \mathrm{D}^{\{f\}}(P \parallel Q) + \mathrm{D}^{\{f\}}(Q \parallel R)$$
$$\mathrm{D}^{\{f\}}(\lambda P_1 + (1-\lambda)P_2 \parallel \lambda Q_1 + (1-\lambda)Q_2) = \lambda\, \mathrm{D}^{\{f\}}(P_1 \parallel Q_1) + (1-\lambda)\, \mathrm{D}^{\{f\}}(P_2 \parallel Q_2).$$

Upper bounding the terms on the right-hand side by $\mathrm{D}^{\mathcal{F}}$ and taking the supremum of the left hand side over $f \in \mathcal{F}$ then gives the claims. The symmetry and positive-definite claims are immediate from the definitions. □

Furthermore, the notion of dual norm has an appealing interpretation in this framework via Remark 2.8, generalizing the fact that total variation distance corresponds to $[0,1]$-valued test functions (or equivalently that ℓ_1 distance corresponds to to $[-1,1]$-valued functions).

Proposition 2.11. *Let $1 \leq p, q \leq \infty$ be Hölder conjugates (meaning $1/p + 1/q = 1$), and let*

$$\mathcal{M}_q \stackrel{\text{def}}{=} \left\{ f : \{0,1\}^m \to \mathbb{R} \;\middle|\; \|f(U_m)\|_q \stackrel{\text{def}}{=} \mathbb{E}[|f(U_m)|^q]^{1/q} \leq 1 \right\}$$

be the set of real-valued functions from $\{0,1\}^m$ with bounded q-th moments. Then $d_{\ell_p} = 2^{-m/q} \cdot d_{\mathcal{M}_q}$, in the sense that for all probability distributions A and B over $\{0,1\}^m$ it holds that $d_{\ell_p}(A,B) = 2^{-m/q} \cdot d_{\mathcal{M}_q}(A,B)$.

In particular, taking $p = 1$ and $q = \infty$ recovers the result for ℓ_1 (equivalently total variation) distance.

Proof. As mentioned this is just the standard fact that the ℓ_p and ℓ_q norms are dual, but for completeness we include a proof in our language using the extremal form of Hölder's inequality (note that since we are dealing with finite probability spaces the extremal equality holds even for $p = \infty$ and $q = 1$). Given probability distributions A and B over $\{0,1\}^m$, we have that

$$\begin{aligned}
d_{\ell_p}(A,B) &= \left(\sum_x |A_x - B_x|^p \right)^{1/p} \\
&= 2^{m/p} \mathop{\mathbb{E}}_{x \sim U_m} [|A_x - B_x|^p]^{1/p} \\
&= 2^{m/p} \max_{\substack{f:\{0,1\}^m \to \mathbb{R} \\ \|f(U_m)\|_q \leq 1}} \left| \mathop{\mathbb{E}}_{x \sim U_m} [f(x)(A_x - B_x)] \right| && \text{(Hölder's extremal equality)} \\
&= 2^{-m+m/p} \max_{\substack{f:\{0,1\}^m \to \mathbb{R} \\ \|f(U_m)\|_q \leq 1}} \Big| \mathbb{E}[f(A)] - \mathbb{E}[f(B)] \Big| \\
&= 2^{-m/q} \cdot d_{\mathcal{M}_q}(A,B) && \text{(by symmetry of } \mathcal{M}_q)
\end{aligned}$$

as desired. □

3 Extractors for weak divergences and connections to samplers

3.1 Definitions

We now use this machinery to extend the notion of an extractor due to Nisan and Zuckerman [NZ96] and the average-case variant of Dodis, Ostrovsky, Reyzin, and Smith [DORS08].

Definition 3.1 (Extends Definition 1.4)**.** Let D be a weak divergence on the set $\{0,1\}^m$, and $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$. Then if for all distributions X over $\{0,1\}^n$ with $\mathrm{H}_\infty(X) \geq k$ it holds that

1. $\mathrm{D}(\mathrm{Ext}(X, U_d) \parallel U_m) \leq \varepsilon$, then Ext is said to be a (k, ε) *extractor for* D, or a (k, ε) D-*extractor.*
2. $\mathbb{E}_{s \sim U_d}[\mathrm{D}(\mathrm{Ext}(X, s) \parallel U_m)] \leq \varepsilon$, then Ext is said to be a (k, ε) *strong extractor for* D, or a (k, ε) *strong* D-*extractor.*

Furthermore, if for all joint distributions (Z, X) where X is distributed over $\{0,1\}^n$ with $\tilde{\mathrm{H}}_\infty(X|Z) \stackrel{\text{def}}{=} \log\bigl(1/\,\mathbb{E}_{z \sim Z}\bigl[2^{-\mathrm{H}_\infty(X|_{Z=z})}\bigr]\bigr) \geq k$, it holds that

3. $\mathbb{E}_{z \sim Z}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, U_d) \parallel U_m) \leq \varepsilon]$, then Ext is said to be a (k, ε) *average-case extractor for* D, or a (k, ε) *average-case* D-*extractor.*
4. $\mathbb{E}_{z \sim Z, s \sim U_d}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, s) \parallel U_m)] \leq \varepsilon$, then Ext is said to be a (k, ε) *average-case strong extractor for* D, or a (k, ε) *average-case strong* D-*extractor.*

Remark 3.2. By taking D to be the total variation distance we recover the standard definitions of extractor and strong extractor due to [NZ96] and the definition of average-case extractor due to [DORS08].

However, our definitions are phrased slightly differently for strong and average-case extractors as an expectation rather than a joint distance, that is, for strong average-case extractors we require a bound on the expectation $\mathbb{E}_{z \sim Z, s \sim U_d}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, s) \parallel U_m)]$ rather than a bound on $\mathrm{D}(Z, U_d, \mathrm{Ext}(X, U_d) \parallel Z, U_d, U_m)$. In our setting, the weak divergence D need not be defined over the larger joint universe, but it is defined for all random variables over $\{0,1\}^m$. In the case of d_{TV} and KL divergence, both definitions are equivalent (for KL divergence, this is an instance of the *chain rule*).

Remark 3.3. The strong variants of Definition 3.1 are also non-strong extractors assuming the weak divergence D is convex in its first argument, as it is for most weak divergences of interest, including the ℓ_p norms for $p \geq 1$, all $\mathrm{D}^{\mathcal{F}}$ defined by test functions, the KL divergence, Rényi divergences for $\alpha \leq 1$, and all Csiszár-Morimoto-Ali-Silvey f-divergences. The average-case variants are always non-average-case extractors by taking Z to be independent of X.

Remark 3.4. We gave Definition 3.1 for general weak divergences which need not be symmetric, and made the particular choice that the output of the extractor was on the left-hand side of the weak divergence and that the uniform distribution was on the right-hand side. This is motivated by the standard information-theoretic divergences such as KL divergence, which require the left-hand distribution to have support contained in the support of the right-hand distribution, and putting the uniform distribution on the right ensures this is always the case. Furthermore, the KL divergence to uniform has a natural interpretation as an entropy difference, $\mathrm{KL}(P \parallel U_m) = m - \mathrm{H}(P)$ for H the Shannon entropy, so that in particular a KL extractor with error ε requires the output to have Shannon entropy at least $m - \varepsilon$. If for a weak divergence D the other direction is more natural, one can always reverse the sides by considering the weak divergence $\mathrm{D}'(Q \parallel P) = \mathrm{D}(P \parallel Q)$.

Remark 3.5. Definition 3.1 does not technically need even a weak divergence, as it suffices to simply have a measure of distance to uniform. However, since weak divergences have minimal constraints, one can define a weak divergence from any distance to uniform by ignoring the second component (or setting it to be infinite for non-uniform distributions).

We also give the natural definition of averaging samplers for arbitrary classes of functions $\mathcal{F}$ extending Definition 1.1, along with the strong variant of Zuckerman [Zuc97].

Definition 3.6. Given a class of functions $\mathcal{F} : \{0,1\}^m \to \mathbb{R}$, a function $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ is said to be a (δ, ε) *strong averaging sampler for* $\mathcal{F}$ or a (δ, ε) *strong averaging* $\mathcal{F}$*-sampler* if for all $f \in \mathcal{F}$, it holds that

$$\Pr_{x \sim U_n}\left[\mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[f_i(\mathrm{Samp}(x)_i) - \mathbb{E}[f_i(U_m)]\Big] > \varepsilon\right] \leq \delta$$

where $[D] = \{1, \ldots, D\}$. If this holds only when $f_1 = \cdots = f_D$, then it is called a *(non-strong)* (δ, ε) *averaging sampler for* $\mathcal{F}$ or (δ, ε) *averaging* $\mathcal{F}$*-sampler.* We say that Samp is a (δ, ε) *strong absolute averaging sampler for* $\mathcal{F}$ if it also holds that

$$\Pr_{x \sim U_n}\left[\left|\mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[f_i(\mathrm{Samp}(x)_i) - \mathbb{E}[f_i(U_m)]\Big]\right| > \varepsilon\right] \leq \delta.$$

with the analogous definition for non-strong samplers.

Remark 3.7. We separated a single-sided version of the error bound in Definition 3.6 as in [Vad12], as it makes the connection between extractors and samplers cleaner and allows us to be specific about what assumptions are needed. Note that if $\mathcal{F}$ is symmetric then every (δ, ε) (strong) sampler for $\mathcal{F}$ is a $(2\delta, \varepsilon)$ (strong) absolute sampler for $\mathcal{F}$, recovering the standard notion up to a factor of 2 in δ.

3.2 Equivalence of extractors and samplers

We now show that Zuckerman's connection [Zuc97] does indeed generalize to this broader setting as promised.

Theorem 3.8. *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be an* $(n - \log(1/\delta), \varepsilon)$*-extractor (respectively strong extractor) for the weak divergence* $\mathrm{D}^{\mathcal{F}}$ *defined by a class of test functions* $\mathcal{F} : \{0,1\}^m \to \mathbb{R}$ *as in Definition 2.5. Then the function* $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ *for* $D = 2^d$ *defined by* $\mathrm{Samp}(x)_i = \mathrm{Ext}(x, i)$ *is a* (δ, ε)*-sampler (respectively strong sampler) for* $\mathcal{F}$*.*

Proof. The proof is essentially the same as that of [Zuc97].

Fix a collection of test functions $f_1, \ldots, f_D \in \mathcal{F}$, where if Ext is not strong we restrict to $f_1 = \cdots = f_D$, and let $B_{f_1,\ldots,f_D} \subseteq \{0,1\}^n$ be defined as

$$\begin{aligned} B_{f_1,\ldots,f_D} &\stackrel{\text{def}}{=} \left\{x \in \{0,1\}^n \;\middle|\; \mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[f_i(\mathrm{Ext}(x,i)) - \mathbb{E}[f_i(U_m)]\Big] > \varepsilon\right\} \\ &= \left\{x \in \{0,1\}^n \;\middle|\; \mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}\big(U_{\{\mathrm{Ext}(x,i)\}} \,\big\|\, U_m\big)\Big] > \varepsilon\right\}, \end{aligned}$$

where $U_{\{z\}}$ is the point mass on z. Then if X is uniform over $B_{f_1,\ldots,f_D}$, we have

$$\begin{aligned} \varepsilon &< \mathop{\mathbb{E}}_{x \sim X}\left[\mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[f_i(\mathrm{Ext}(x,i)) - \mathbb{E}[f_i(U_m)]\Big]\right] \\ &= \mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X,i) \,\|\, U_m)\Big] \\ &= \begin{cases} \mathrm{D}^{\{f_1\}}(\mathrm{Ext}(X,U_d) \,\|\, U_m) & \text{if } f_1 = \cdots = f_D \\ \mathbb{E}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X,i) \,\|\, U_m)\Big] & \text{always} \end{cases} \\ &\leq \begin{cases} \mathrm{D}^{\mathcal{F}}(\mathrm{Ext}(X,U_d) \,\|\, U_m) & \text{if } f_1 = \cdots = f_D \\ \mathbb{E}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\mathcal{F}}(\mathrm{Ext}(X,i) \,\|\, U_m)\Big] & \text{always} \end{cases} \end{aligned}$$

Since Ext is an $(n - \log(1/\delta), \varepsilon)$-extractor (respectively strong extractor) for $\mathrm{D}^{\mathcal{F}}$ we must have $\mathrm{H}_\infty(X) < n - \log(1/\delta)$. But $\mathrm{H}_\infty(X) = \log|B_{f_1,\ldots,f_D}|$ by definition, so we have $|B_{f_1,\ldots,f_D}| < \delta 2^n$. Hence, the probability that a random $x \in \{0,1\}^n$ lands in $B_{f_1,\ldots,f_D}$ is less than δ, and since $B_{f_1,\ldots,f_D}$ is exactly the set of seeds which are bad for Samp, this concludes the proof. ∎

Remark 3.9. Hölder's inequality implies that an extractor for ℓ_p with error $\varepsilon \cdot 2^{-m(p-1)/p}$ is also an ℓ_1 extractor and thus $[-1,1]$-averaging sampler with error ε. Proposition 2.11 and Theorem 3.8 show that they are in fact samplers for the much larger class of functions $\mathcal{M}_{p/(p-1)}$ with bounded $p/(p-1)$ moments (rather than just ∞ moments), also with error ε.

Furthermore, if all the functions in $\mathcal{F}$ have bounded deviation from their mean (for example, subgaussian functions from $f : \{0,1\}^m \to \mathbb{R}$ have such a bound of $O(\sqrt{m})$ by the tail bounds from Lemma 4.3), then we also have a partial converse that recovers the standard converse in the case of total variation distance.

Theorem 3.10. *Let $\mathcal{F}$ be a class of functions $\mathcal{F} \subset \{0,1\}^m \to \mathbb{R}$ with finite* maximum deviation from the mean, *meaning* $\max\operatorname{dev}(\mathcal{F}) \stackrel{\text{def}}{=} \sup_{f\in\mathcal{F}} \max_{x\in\{0,1\}^n}\big(f(x) - \mathbb{E}[f(U_m)]\big) < \infty$*. Then given a (δ,ε) $\mathcal{F}$-sampler (respectively (δ,ε) strong $\mathcal{F}$-sampler)* $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$*, the function* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for $d = \log D$ defined by* $\mathrm{Ext}(x,i) = \mathrm{Samp}(x)_i$ *is a $\big(k, \varepsilon + \delta \cdot 2^{n-k} \cdot \max\operatorname{dev}(\mathcal{F})\big)$ $\mathrm{D}^{\mathcal{F}}$-extractor (respectively strong $\mathrm{D}^{\mathcal{F}}$-extractor) for every $0 \le k \le n$.*

In particular, Ext *is an $\big(n - \log(1/\delta) + \log(1/\eta), \varepsilon + \eta \cdot \max\operatorname{dev}(\mathcal{F})\big)$ average-case $\mathrm{D}^{\mathcal{F}}$-extractor (respectively strong average-case $\mathrm{D}^{\mathcal{F}}$-extractor) for every $\delta \le \eta \le 1$.*

Proof. Again the proof is analogous to the one in [Zuc97].

Fix a distribution X over $\{0,1\}^m$ with $\mathrm{H}_\infty(X) \ge k$ and a collection of test functions $f_1, \dots, f_D \in \mathcal{F}$, where if Samp is not strong we restrict to $f_1 = \cdots = f_D$. Then since Samp is a (δ,ε) $\mathcal{F}$-sampler, we know that the set of seeds for which the sampler is bad must be small. Formally, the set

$$\begin{aligned} B_{f_1,\dots,f_D} &\stackrel{\text{def}}{=} \left\{ x \in \{0,1\}^n \,\middle|\, \mathop{\mathbb{E}}_{i\sim U_d}\Big[f_i(\mathrm{Samp}(x)_i) - \mathbb{E}[f_i(U_m)]\Big] > \varepsilon \right\} \\ &= \left\{ x \in \{0,1\}^n \,\middle|\, \mathop{\mathbb{E}}_{i\sim U_d}\Big[f_i(\mathrm{Ext}(x,i)) - \mathbb{E}[f_i(U_m)]\Big] > \varepsilon \right\} \end{aligned}$$

has size $|B_{f_1,\dots,f_D}| \le \delta 2^n$. Thus, since X has min-entropy at least k we know $\Pr[X \in B_{f_1,\dots,f_D}] \le 2^{-k} \cdot \delta 2^n$, so we have

$$\begin{aligned} &\mathop{\mathbb{E}}_{i\sim U_d}\Big[\mathbb{E}\Big[f_i(\mathrm{Ext}(X,i)) - \mathbb{E}[f_i(U_m)]\Big]\Big] \\ &\qquad = \mathop{\mathbb{E}}_{X}\Big[\mathop{\mathbb{E}}_{i\sim U_d}\Big[f_i(\mathrm{Ext}(X,i)) - \mathbb{E}[f_i(U_m)]\Big]\Big] \\ &\qquad = \Pr[X \in B_{f_1,\dots,f_D}] \cdot \mathop{\mathbb{E}}_{X}\left[\mathop{\mathbb{E}}_{i\sim U_d}\Big[f_i(\mathrm{Ext}(X,i)) - \mathbb{E}[f_i(U_m)]\Big] \,\middle|\, X \in B_{f_1,\dots,f_D}\right] \\ &\qquad\quad + \Pr[X \notin B_{f_1,\dots,f_D}] \cdot \mathop{\mathbb{E}}_{X}\left[\mathop{\mathbb{E}}_{i\sim U_d}\Big[f_i(\mathrm{Ext}(X,i)) - \mathbb{E}[f_i(U_m)]\Big] \,\middle|\, X \notin B_{f_1,\dots,f_D}\right] \\ &\qquad \le \Pr[X \in B_{f_1,\dots,f_D}] \cdot \max\operatorname{dev}(\mathcal{F}) + \Pr[X \notin B_{f_1,\dots,f_D}] \cdot \varepsilon \\ &\qquad \le 2^{-k} \cdot \delta 2^n \cdot \max\operatorname{dev}(\mathcal{F}) + \varepsilon \end{aligned}$$

completing the proof of the main claim. The "in particular" statement follows since if (Z,X) are jointly distributed with $\tilde{\mathrm{H}}_\infty(X|Z) \ge n - \log(1/\delta) + \log(1/\eta)$ we have

$$\mathop{\mathbb{E}}_{z\sim Z}\Big[\varepsilon + \delta \cdot 2^{n - \mathrm{H}_\infty(X|_{Z=z})} \cdot \max\operatorname{dev}(\mathcal{F})\Big] = \varepsilon + \delta \cdot 2^{n - \tilde{\mathrm{H}}_\infty(X|Z)} \cdot \max\operatorname{dev}(\mathcal{F}) \le \varepsilon + \eta \cdot \max\operatorname{dev}(\mathcal{F})$$

by definition of conditional min-entropy. □

3.3 All extractors are average-case

Under a similar boundedness condition for general weak divergences, we can recover the standard fact that all extractors are average-case extractors under a slight loss of parameters (the same loss as achieved by

Dodis, Ostrovsky, Reyzin, and Smith [DORS08] for the case of total variation distance). More interestingly, if the weak divergence is given by $\mathrm{D}^{\mathcal{F}}$ for a symmetric class of (possibly unbounded) functions $\mathcal{F}$, we can also generalize and recover the result of Vadhan [Vad12, Problem 6.8] that shows that a (k, ε) extractor (for total variation) is a $(k, 3\varepsilon)$ average-case extractor without any other loss.

Theorem 3.11. *Let* D *be a bounded weak divergence over* $\{0,1\}^m$*, meaning that*

$$0 \le \|\mathrm{D}\|_\infty \stackrel{\text{def}}{=} \sup_{P \text{ on } \{0,1\}^m} \mathrm{D}(P \parallel U_m) < \infty.$$

Then a (k, ε)*-extractor for* D *(respectively strong extractor)* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is also a* $(k + \log(1/\eta), \varepsilon + \eta \cdot \|\mathrm{D}\|_\infty)$ *average-case-extractor for* D *(respectively strong average-case-extractor) for any* $0 < \eta \le 1$.

Proof. The proof is analogous to that of [DORS08]. We prove it only for non-strong extractors, the proof for strong extractors is completely analogous by adding more expectations.

For jointly distributed random variables (Z, X) such that $\tilde{\mathrm{H}}_\infty(X|Z) \ge k + \log(1/\eta)$, we have by [DORS08, Lemma 2.2] that the probability that $\Pr_{z\sim Z}[\mathrm{H}_\infty(X|_{Z=z}) < k] \le \eta$. Thus

$$\begin{aligned}
&\mathop{\mathbb{E}}_{z\sim Z}\big[\mathrm{D}\big(\mathrm{Ext}\big(X|_{Z=z}, U_d\big) \parallel U_m\big)\big] \\
&\quad= \Pr_{z\sim Z}[\mathrm{H}_\infty(X|_{Z=z}) < k] \cdot \mathop{\mathbb{E}}_{z\sim Z}\big[\mathrm{D}\big(\mathrm{Ext}\big(X|_{Z=z}, U_d\big) \parallel U_m\big) \,\big|\, \mathrm{H}_\infty(X|_{Z=z}) < k\big] \\
&\qquad+ \Pr_{z\sim Z}[\mathrm{H}_\infty(X|_{Z=z}) \ge k] \cdot \mathop{\mathbb{E}}_{z\sim Z}\big[\mathrm{D}\big(\mathrm{Ext}\big(X|_{Z=z}, U_d\big) \parallel U_m\big) \,\big|\, \mathrm{H}_\infty(X|_{Z=z}) \ge k\big] \\
&\quad\le \eta \cdot \|D\|_\infty + 1 \cdot \varepsilon
\end{aligned}$$

□

Theorem 3.12. *Let* $\mathcal{F}$ *be a symmetric class of test functions and* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a* (k, ε) *extractor (respectively strong extractor) for* $\mathrm{D}^{\mathcal{F}}$*, where* k *is at most* $n-1$*. Then* Ext *is an* $(k, 3\varepsilon)$ *average-case extractor (respectively strong average-case extractor) for* $\mathrm{D}^{\mathcal{F}}$*.*

Remark 3.13. Theorem 3.12 also applies to extractors for the ℓ_p norms via Proposition 2.11.

The proof of Theorem 3.12 follows the strategy outlined by Vadhan [Vad12, Problem 6.8]. We first isolate the following key lemma which shows that any extractor with error that gracefully decays with lower min-entropy is average-case with minimal loss of parameters, as opposed to Theorem 3.11 which used a worst-case error bound when the min-entropy is low.

Lemma 3.14. *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a* (k, ε) *extractor (respectively strong extractor) for* D *such that for every* $0 \le t \le k$, Ext *is also a* $(k - t, 2^{t+1} \cdot \varepsilon)$ *extractor (respectively strong extractor) for* D*. Then* Ext *is a* $(k, 3\varepsilon)$ *average-case extractor (respectively strong average-case extractor) for* D*.*

Proof. We prove this for strong extractors, the non-strong case is analogous. For every (Z, X) with X distributed on $\{0,1\}^n$ and $\tilde{\mathrm{H}}_\infty(X|Z) \ge k$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{z\sim Z, s\sim U_d}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, s) \parallel U_m)] &= \mathop{\mathbb{E}}_{z\sim Z}\left[\mathop{\mathbb{E}}_{s\sim U_d}[\mathrm{D}(\mathrm{Ext}(X|_{Z=z}, s) \parallel U_m)]\right] \\
&\le \mathop{\mathbb{E}}_{z\sim Z}\left[\begin{cases} \varepsilon & \text{if } \mathrm{H}_\infty(X|_{Z=z}) \ge k \\ 2^{k - \mathrm{H}_\infty(X|_{Z=z}) + 1} \cdot \varepsilon & \text{otherwise} \end{cases}\right] \\
&\le \varepsilon \cdot \mathop{\mathbb{E}}_{z\sim Z}\left[1 + 2^{k - \mathrm{H}_\infty(X|_{Z=z}) + 1}\right] \le 3\varepsilon
\end{aligned}$$

where the last inequality follows from the fact that $\mathbb{E}_{z\sim Z}\big[2^{-\mathrm{H}_\infty(X|_{Z=z})}\big] = 2^{-\tilde{\mathrm{H}}_\infty(X|Z)}$ by definition of conditional min-entropy. □

Proof of Theorem 3.12. By the previous lemma, it suffices to prove that for every $t \geq 0$, Ext is a $(k - t, (2^{t+1} - 1) \cdot \varepsilon)$ extractor (respectively strong extractor) for $\mathrm{D}^{\mathcal{F}}$. Since $\mathrm{D}^{\mathcal{F}}$ is convex in its first argument by Lemma 2.10, following Chor and Goldreich [CG88] it is enough to consider only distributions with min-entropy $k - t$ that are supported on a set of at most 2^{n-1}. Fix such a distribution X and a collection of test functions $f_1, \ldots, f_D \in \mathcal{F}$ with $f_1 = \cdots = f_D$ if Ext is not strong. Then since X is supported on a set of size at most 2^{n-1}, the distribution Y that is uniform over the complement of $\mathrm{Supp}(X)$ has min-entropy at least $n - 1 \geq k$, and furthermore the mixture $2^{-t}X + (1 - 2^{-t})Y$ has min-entropy at least k. Hence, as Ext is a (k, ε) extractor (respectively strong extractor) for $\mathrm{D}^{\mathcal{F}}$,

$$
\begin{aligned}
\varepsilon &\geq \mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}\big(\mathrm{Ext}\big(2^{-t}X + (1 - 2^{-t})Y, i\big) \,\|\, U_m\big)\Big] \\
&= 2^{-t} \mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X, i) \,\|\, U_m)\Big] + (1 - 2^{-t}) \mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(Y, i) \,\|\, U_m)\Big] \\
&= 2^{-t} \mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X, i) \,\|\, U_m)\Big] - (1 - 2^{-t}) \mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\{c_i - f_i\}}(\mathrm{Ext}(Y, i) \,\|\, U_m)\Big] \\
&\geq 2^{-t} \mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X, i) \,\|\, U_m)\Big] - (1 - 2^{-t}) \cdot \varepsilon \qquad \text{(since } \mathrm{H}_\infty(Y) \geq k\text{)} \\
\big(2^{t+1} - 1\big) \cdot \varepsilon &\geq \mathop{\mathbb{E}}_{i \sim U_{[D]}}\Big[\mathrm{D}^{\{f_i\}}(\mathrm{Ext}(X, i) \,\|\, U_m)\Big]
\end{aligned}
$$

where $c_i \in \mathbb{R}$ is such that $c_i - f_i \in \mathcal{F}$ as guaranteed to exist by the symmetry of $\mathcal{F}$. □

4 Subgaussian distance and connections to other notions

Now that we've introduced the general machinery we need, we can go back to our motivation of subgaussian samplers. We will need some standard facts about subgaussian and subexponential random variables, we recommend the book of Vershynin [Ver18] for an introduction.

Definition 4.1. A real-valued mean-zero random variable Z is said to be *subgaussian with parameter σ* if for every $t \in \mathbb{R}$ the moment generating function of Z is bounded as

$$\ln \mathbb{E}\big[e^{tZ}\big] \leq \frac{t^2\sigma^2}{2}.$$

If this is only holds for $|t| \leq b$ then Z is said to be *(σ, b)-subgamma*, and if Z is $(\sigma, 1/\sigma)$-subgamma then Z is said to be *subexponential with parameter σ*.

Remark 4.2. There are many definitions of subgaussian (and especially subexponential) random variables in the literature, but they are all equivalent up to constant factors in σ and only affect constants already hidden in big-O's.

Lemma 4.3. *Let Z be a real-valued random variable. Then*

1. *(Hoeffding's lemma) If Z is bounded in the interval $[0, 1]$, then $Z - \mathbb{E}[Z]$ is subgaussian with parameter $1/2$.*
2. *If Z is mean-zero, then Z is subgaussian (respectively subexponential) with parameter σ if and only if cZ is subgaussian (respectively subexponential) with parameter $|c|\sigma$ for every $c \neq 0$.*

Furthermore, if Z is mean-zero and subgaussian with parameter σ, then

1. *For all $t > 0$, $\max\big(\Pr[Z > t], \Pr[Z < -t]\big) \leq e^{-t^2/2\sigma^2}$.*
2. *$\|Z\|_p \stackrel{\text{def}}{=} \mathbb{E}[|Z|^p]^{1/p} \leq 2\sigma\sqrt{p}$ for all $p \geq 1$.*

3. Z is subexponential with parameter σ.

We are now in a position to formally define the *subgaussian distance*.

Definition 4.4. For every finite set $\mathcal{X}$, we define the set $\mathcal{G}_{\mathcal{X}}$ of *subgaussian test functions on $\mathcal{X}$* (respectively the set $\mathcal{E}_{\mathcal{X}}$ of *subexponential test functions on $\mathcal{X}$*) to be the set of functions $f : \mathcal{X} \to \mathbb{R}$ such that the random variable $f(U_{\mathcal{X}})$ is mean-zero and subgaussian (respectively subexponential) with parameter $1/2$. Then $\mathcal{G}_{\mathcal{X}}$ and $\mathcal{E}_{\mathcal{X}}$ are symmetric and distinguishing, so by Lemma 2.10 the respective distances induced by $\mathcal{G}_{\mathcal{X}}$ and $\mathcal{E}_{\mathcal{X}}$ are jointly convex metrics called the *subgaussian distance* and *subexponential distance* respectively and are denoted as $d_{\mathcal{G}}(P,Q)$ and $d_{\mathcal{E}}(P,Q)$.

Remark 4.5. We choose subgaussian parameter $1/2$ in Definition 4.4 as by Hoeffding's lemma, all functions $f : \{0,1\}^m \to [0,1]$ have that $f(U_m) - \mathbb{E}[f(U_m)]$ is subgaussian with parameter $1/2$, so this choice preserves the same "scale" as total variation distance. However, the choice of parameter is essentially irrelevant by linearity, as different choices of parameter simply scale the metric $d_{\mathcal{G}}$.

Note that absolute averaging samplers for $\mathcal{G}_{\{0,1\}^m}$ from Definition 3.6 are exactly subgaussian samplers as defined in the introduction. Thus, by Remark 3.7 and Theorem 3.8, to construct subgaussian samplers it is enough to construct extractors for the subgaussian distance $d_{\mathcal{G}}$.

4.1 Composition

Unfortunately, the subgaussian distance has a major disadvantage compared to total variation distance that complicates extractor construction: it does not satisfy the data-processing inequality, that is, there are probability distributions P and Q over a set A and a function $f : A \to B$ such that

$$d_{\mathcal{G}}(f(P), f(Q)) \not\leq d_{\mathcal{G}}(P,Q).$$

This happens because subgaussian distance is defined by functions which are required to be subgaussian only with respect to the *uniform distribution*. A simple explicit counterexample comes from taking $f : \{0,1\}^1 \to \{0,1\}^m$ defined by $x \mapsto (x, 0^{m-1})$ and taking P to be the point mass on 0 and Q the point mass on 1. Their subgaussian distance in $\{0,1\}^1$ is obviously $O(1)$, but the subgaussian distance of $f(P)$ and $f(Q)$ in $\{0,1\}^m$ is $\Theta(\sqrt{m})$.

The reason this matters because a standard operation (cf. Nisan and Zuckerman [NZ96]; Goldreich and Wigderson [GW97]; Reingold, Vadhan, and Wigderson [RVW00]) in the construction of samplers and extractors for bounded functions is to do the following: given extractors

$$\mathrm{Ext}_{out} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$$
$$\mathrm{Ext}_{in} : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^d,$$

define $\mathrm{Ext} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \to \{0,1\}^m$ by

$$\mathrm{Ext}\big((x,y),s\big) = \mathrm{Ext}_{out}\big(x, \mathrm{Ext}_{in}(y,s)\big).$$

The reason this works for total variation distance is exactly the data-processing inequality: if Y has enough min-entropy given X, then $\mathrm{Ext}_{in}(Y, U_{d'})$ will be close in total variation distance to U_d, and by the data-processing inequality for total variation distance this closeness is not lost under the application of Ext_{out}. The assumption that Y has min-entropy given X means that (X,Y) is a so-called *block-source*, and is implied by (X,Y) having enough min-entropy as a joint distribution. From the sampler perspective, this construction uses the inner sampler Ext_{in} to subsample the outer sampler. On the other hand, for subgaussian distance, the distribution $\mathrm{Ext}_{in}(Y, U_{d'})$ can be ε-close to uniform but still have some element with excess probability mass $\Omega(\varepsilon/\sqrt{d})$, and this element (seed) when mapped by Ext_{out} can retain[2] this excess mass in $\{0,1\}^m$,

[2]Given a subgaussian extractor Ext with $d \geq \log(m/\varepsilon)$, adding a single extra seed $*$ to Ext such that $\mathrm{Ext}(x,*) = 0^m$ results in a subgaussian extractor with error at most $2^{-d} \cdot \sqrt{2m} + \varepsilon \leq 3\varepsilon$ by convexity of $d_{\mathcal{G}}$ and the fact that $\left\| d_{\mathcal{G}_{\{0,1\}^m}} \right\|_\infty < \sqrt{2m}$.

which results in subgaussian distance $\Theta(\varepsilon\sqrt{m/d}) \gg \varepsilon$. Similarly, from the sampler perspective, even when the outer sampler Ext_{out} is a good subgaussian sampler for $\{0,1\}^m$, there is no reason that a good subgaussian sampler Ext_{in} for $\{0,1\}^d$ the seeds of Ext_{out} will preserve the larger sampler property when $m \gg d$.

Thus, since this composition operation is needed to construct high-min entropy extractors with the desired seed length even for total variation distance, to construct such extractors for subgaussian distance we need to bypass this barrier. The natural approach is to construct extractors for a better-behaved weak divergence that bounds the subgaussian distance.

Remark 4.6. Similar reasoning shows that if Ext is a strong (k,ε) subgaussian extractor, then it is not necessarily the case that the function $(x,s) \mapsto (s, \mathrm{Ext}(x,s))$ that prepends the seed to the output is a (non-strong) (k,ε) subgaussian extractor (in contrast to extractors for total variation distance), though the converse does hold.

4.2 Connections to other weak divergences

Therefore, to aid in extractor construction, we show how $d_{\mathcal{G}}$ relates to other statistical weak divergences.

Most basically, the subgaussian distance over $\{0,1\}^m$ differs from total variation distance up to a factor of $O(\sqrt{m})$.

Lemma 4.7. *Let P and Q be distributions on $\{0,1\}^m$. Then*

$$d_{TV}(P,Q) \le d_{\mathcal{G}}(P,Q) \le \sqrt{2\ln 2 \cdot m} \cdot d_{TV}(P,Q)$$

Proof. That $d_{TV} \le d_{\mathcal{G}}$ is immediate from Hoeffding's lemma and the discussion in Remark 4.5. The reverse bound holds since any subgaussian function takes values at most $\sqrt{\ln 2/2 \cdot m}$ away from the mean by the tail bounds from part 3 of Lemma 4.3, and so any subgaussian test function f has the property that $1/2 + f/\sqrt{2\ln 2 \cdot m}$ is $[0,1]$-valued and thus lower bounds the total variation distance. □

While this allows constructing subgaussian extractors and samplers from total variation extractors, as discussed in the introduction the fact that the upper bound depends on m leads to suboptimal bounds. By starting with a stronger measure of error, we pay a much smaller penalty.

Lemma 4.8. *Let P and Q be distributions on $\{0,1\}^m$. Then for every $\alpha > 0$*

$$2d_{TV}(P,Q) = d_{\ell_1}(P,Q) \le 2^{m\alpha/(1+\alpha)} \cdot d_{\ell_{1+\alpha}}(P,Q)$$
$$d_{\mathcal{G}}(P,Q) \le 2^{m\alpha/(1+\alpha)}\sqrt{1+\frac{1}{\alpha}} \cdot d_{\ell_{1+\alpha}}(P,Q)$$

In particular, that there is only an additional $\sqrt{1+1/\alpha}$ factor when moving to subgaussian distance compared to total variation, which in particular does not depend on m and is constant for constant α.

Proof. By Proposition 2.11, for any function $f : \{0,1\}^m \to \mathbb{R}$ it holds that

$$\mathrm{D}^{\{f\}}(P \parallel Q) \le \|f(U_m)\|_{1+\frac{1}{\alpha}} \cdot d_{\mathcal{M}_{1+\frac{1}{\alpha}}}(P,Q) = \|f(U_m)\|_{1+\frac{1}{\alpha}} \cdot 2^{m\alpha/(1+\alpha)} \cdot d_{\ell_{1+\alpha}}(P,Q).$$

The result follows since $[-1,1]$-valued functions f satisfy moment bounds $\|f(U_m)\|_q \le 1$ for all $q \ge 1$, and functions f which are subgaussian satisfy moment bounds $\|f(U_m)\|_q \le \sqrt{q}$ by Lemma 4.3. □

One downside of starting with bounds on $\ell_{1+\alpha}$ is that, extending a well-known linear seed length linear bound for ℓ_2-extractors (e.g. [Vad12, Problem 6.4]), we show in Corollary 5.29 that for every $1 > \alpha > 0$, there is a constant $c_\alpha > 0$ such any $\ell_{1+\alpha}$ extractor with error smaller than $c_\alpha \cdot 2^{-m\alpha/(1+\alpha)}$ requires seed length linear in $\alpha \cdot \min(n-k, m)$, for $n-k$ the entropy deficiency and m the output length. One might hope that sending α to 0 would eliminate this linear lower bound but still bound the subgaussian distance, but phrased this way sending α to 0 just results in a total variation extractor.

However, with a shift in perspective essentially the same approach works: by Example 2.4, $d_{\ell_2}(P, U_m) \leq \varepsilon \cdot 2^{-m/2}$ implies $\mathrm{D}_2(P \parallel U_m) \leq \varepsilon^2/\ln 2$, and there is an analogous linear seed length lower bound on constant error $\mathrm{D}_{1+\alpha}$ extractors for every $\alpha > 0$. In this case, however, sending α to 0 results in the *KL divergence*, which does upper bound the subgaussian distance, and in fact with the same parameters as for total variation distance.

Lemma 4.9. *Let P be a distribution on $\{0,1\}^m$. Then*

$$d_{\mathcal{G}}(P, U_m) \leq \sqrt{\frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m)}$$
$$d_{\mathcal{E}}(P, U_m) \leq \begin{cases} \sqrt{\frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m)} & \text{if } \mathrm{KL}(P \parallel U_m) \leq \frac{1}{2\ln 2} \\ \frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m) + \frac{1}{4} & \text{if } \mathrm{KL}(P \parallel U_m) > \frac{1}{2\ln 2} \end{cases}$$

where these bounds are concave in $\mathrm{KL}(P \parallel U_m)$*. In the reverse direction, it holds that*

$$\mathrm{KL}(P \parallel U_m) \leq m \cdot d_{TV}(P, U_m) + h(d_{TV}(P, U_m))$$

where $h(x) = x \log(1/x) + (1-x)\log(1/(1-x))$ *is the (concave) binary entropy function.*

Proof. The upper bound on subgaussian distance follows from a general form of Pinsker's inequality as in [BLM13, Lemma 4.18], but for the extension to subexponential functions we reproduce its proof here, based on the Donsker–Varadhan "variational" formulation of KL divergence [DV76] (cf. [BLM13, Corollary 4.15])

$$\mathrm{KL}(P \parallel U_m) = \frac{1}{\ln 2} \cdot \sup_{g:\{0,1\}^m \to \mathbb{R}} \Big(\mathbb{E}[g(P)] - \ln \mathbb{E}\big[e^{g(U_m)}\big] \Big).$$

Now if $f : \{0,1\}^m \to \mathbb{R}$ satisfies $\mathbb{E}[f(U_m)] = 0$, then by letting $g(x) = t \cdot f(x)$, this implies

$$\mathbb{E}[f(P)] - \mathbb{E}[f(U_m)] = \frac{1}{t} \cdot \mathbb{E}[g(P)] \leq \frac{\ln 2 \cdot \mathrm{KL}(P \parallel U_m) + \ln \mathbb{E}\big[e^{t \cdot f(U_m)}\big]}{t}$$

for all $t > 0$. Thus, when $\ln \mathbb{E}\big[e^{t\cdot f(U_m)}\big] \leq t^2/8$, we have $\mathbb{E}[f(P)] - \mathbb{E}[f(U_m)] \leq \ln 2 \cdot \mathrm{KL}(P \parallel U_m)/t + t/8$.

Then since subgaussian random variables satisfy such a bound for all t, we can make the optimal choice $t = \sqrt{8 \ln 2 \cdot \mathrm{KL}(P \parallel U_m)}$ to get the claimed bound on $d_{\mathcal{G}}$. For subexponential random variables, which satisfy such a bound only for $|t| \leq 2$, we choose $t = \min(\sqrt{8 \ln 2 \cdot \mathrm{KL}(P \parallel U_m)}, 2)$, which gives

$$d_{\mathcal{E}}(P, U_m) \leq \begin{cases} \sqrt{\frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m)} & \text{if } \mathrm{KL}(P \parallel U_m) \leq \frac{1}{2\ln 2} \\ \frac{\ln 2}{2} \cdot \mathrm{KL}(P \parallel U_m) + \frac{1}{4} & \text{if } \mathrm{KL}(P \parallel U_m) > \frac{1}{2\ln 2} \end{cases}$$

as desired. The concavity of this bound follows by noting that it has a continuous and nonincreasing derivative.

For the reverse inequality, we use a bound on the difference in entropy between distributions P and Q on a set of size S which states

$$|H(P) - H(Q)| \leq \lg(S-1) \cdot d_{TV}(P, Q) + h(d_{TV}(P, Q)).$$

This inequality is a simple consequence of Fano's inequality as noted by Goldreich and Vadhan [GV99, Fact B.1], and implies the desired result by taking $Q = U_m$ as $\mathrm{KL}(P \parallel U_m) = H(U_m) - H(P)$ and $|\{0,1\}^m| = 2^m$. □

Remark 4.10. There are sharper upper bounds on the KL divergence than given in Lemma 4.9, such as the bound of Audenaert and Eisert [AE05, Theorem 6], but the bound we use has the advantage of being defined for the entire range of the total variation distance and being everywhere concave.

5 Extractors for KL divergence

By Lemma 4.9, the subgaussian distance can be bounded in terms of the KL divergence to uniform, so by the following easy lemma to construct subgaussian extractors it suffices to construct extractors for KL divergence.

Lemma 5.1. *Let V_1 and V_2 be weak divergences on the set $\{0,1\}^m$ and $f : \mathbb{R} \to \mathbb{R}$ be a function such that $V_1(P \parallel U_M) \leq f(V_2(P \parallel U_m))$ for all distributions P on $\{0,1\}^m$. Then if f is increasing on $(0, \varepsilon)$, every (k, ε) extractor Ext for V_1 is also a $(k, f(\varepsilon))$-extractor for V_2, and if f is also concave, then if Ext is strong or average-case as a V_1-extractor, it has the same properties as a $(k, f(\varepsilon))$ extractor for V_2.*

Importantly, the KL divergence does not have the flaws of subgaussian distance discussed in Section 4.1. The classic *data-processing inequality* says that KL divergence is non-increasing under postprocessing by (possibly randomized) functions, and the *chain rule* for KL divergence says that

$$\mathrm{KL}(A, B \parallel X, Y) = \mathrm{KL}(A \parallel X) + \mathop{\mathbb{E}}_{a \sim A}[\mathrm{KL}(B|_{A=a} \parallel Y|_{X=a})]$$

for all distributions A, B, X, and Y, so that in particular

$$\mathop{\mathbb{E}}_{s \sim U_d}[\mathrm{KL}(\mathrm{Ext}(X, s) \parallel U_m)] = \mathrm{KL}(U_d, \mathrm{Ext}(X, U_d) \parallel U_d, U_m)$$

and prepending the seed of a strong KL-extractor does in fact give a non-strong KL-extractor:

Lemma 5.2. *A function $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a (k, ε) strong KL-extractor (respectively strong average-case KL-extractor) if and only if the function $\mathrm{Ext}' : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^{d+m}$ defined by $\mathrm{Ext}'(x, s) = (s, \mathrm{Ext}(x, s))$ is a (non-strong) (k, ε) KL-extractor (respectively average-case KL-extractor).*

Furthermore, KL divergence satisfies a type of triangle inequality when combined with higher Rényi divergences:

Lemma 5.3. *Let P, Q, and R be distributions over a finite set $\mathcal{X}$. Then for all $\alpha > 0$, it holds that*

$$\mathrm{KL}(P \parallel R) \leq \left(1 + \frac{1}{\alpha}\right) \cdot \mathrm{KL}(P \parallel Q) + \mathrm{D}_{1+\alpha}(Q \parallel R)$$

Proof. This follows from a characterization of Rényi divergence due to van Erven and Harremoës [vE10, Lemma 6.6] [vEH14, Theorem 30] and Shayevitz [Sha11, Theorem 1], who prove that for for every positive real $\beta \neq 1$ and distributions X and Y that

$$(1 - \beta)\,\mathrm{D}_\beta(X \parallel Y) = \inf_Z \bigl\{\beta\,\mathrm{KL}(Z \parallel X) + (1 - \beta)\,\mathrm{KL}(Z \parallel Y)\bigr\}.$$

In particular, choosing $\beta = 1 + \alpha$, $X = Q$, and $Y = R$ and upper bounding the infimum by the particular choice of $Z = P$ gives the claim. ☐

5.1 Composition

These properties imply that composition does work as we want (without any loss depending on the output length m) assuming we have extractors for KL and higher divergences.

Theorem 5.4 (Composition for high min-entropy Rényi entropy extractors, cf. [GW97])**.** *Suppose*

1. $\mathrm{Ext}_{out} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is an* $(n - \log(1/\delta), \varepsilon_{out})$ *extractor for* $\mathrm{D}_{1+\alpha}$ *with* $\alpha > 0$*,*
2. $\mathrm{Ext}_{in} : \{0,1\}^{n'} \times \{0,1\}^{d'} \to \{0,1\}^d$ *is an* $(n' - \log(1/\delta), \varepsilon_{in})$ *average-case KL-extractor,*

and define $\mathrm{Ext} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \to \{0,1\}^m$ by $\mathrm{Ext}\bigl((x, y), s\bigr) = \mathrm{Ext}_{out}(x, \mathrm{Ext}_{in}(y, s))$. Then Ext is an $(n + n' - \log(1/\delta), \varepsilon_{out} + (1 + 1/\alpha) \cdot \varepsilon_{in})$ extractor for KL. Furthermore, if Ext_{in} is a strong average-case KL-extractor, then Ext is a strong KL-extractor, and if Ext_{out} is average-case then so is Ext.

Proof. Let (Z, X, Y) be jointly distributed random variables with X distributed over $\{0,1\}^n$ and Y over $\{0,1\}^{n'}$ such that $\tilde{\mathrm{H}}_\infty(X, Y|Z) \geq n + n' - \log(1/\delta)$. Let S' be a distribution over $\{0,1\}^{d'}$ which is independent of X, Y, and Z. Then for every $z \in \mathrm{Supp}(Z)$, we have by Lemma 5.3 and the data-processing inequality for KL divergence that

$$\begin{aligned}
&\mathrm{KL}(\mathrm{Ext}((X|_{Z=z}, Y|_{Z=z}), S') \parallel U_m) \\
&\quad = \mathrm{KL}(\mathrm{Ext}_{out}(X|_{Z=z}, \mathrm{Ext}_{in}(Y|_{Z=z}, S')) \parallel U_m) \\
&\quad \leq (1 + 1/\alpha) \cdot \mathrm{KL}(\mathrm{Ext}_{out}(X|_{Z=z}, \mathrm{Ext}_{in}(Y|_{Z=z}, S')) \parallel \mathrm{Ext}_{out}(X|_{Z=z}, U_d)) \\
&\quad\qquad + \mathrm{D}_{1+\alpha}(\mathrm{Ext}_{out}(X|_{Z=z}, U_d) \parallel U_m) \\
&\quad \leq (1 + 1/\alpha) \cdot \mathrm{KL}(X|_{Z=z}, \mathrm{Ext}_{in}(Y|_{Z=z}, S') \parallel X|_{Z=z}, U_d) + \mathrm{D}_{1+\alpha}(\mathrm{Ext}_{out}(X|_{Z=z}, U_d) \parallel U_m) \\
&\quad = (1 + 1/\alpha) \cdot \mathop{\mathbb{E}}_{x \sim X|_{Z=z}} [\mathrm{KL}(\mathrm{Ext}_{in}(Y|_{X=x, Z=z}, S') \parallel U_d)] + \mathrm{D}_{1+\alpha}(\mathrm{Ext}_{out}(X|_{Z=z}, U_d) \parallel U_m)
\end{aligned}$$

where the last equality follows from the chain rule for KL divergence. Now by standard properties of conditional min-entropy (see for example [DORS08, Lemma 2.2]), we know that $\tilde{\mathrm{H}}_\infty(X|Z) \geq \tilde{\mathrm{H}}_\infty(X, Y|Z) - \log|\mathrm{Supp}(Y)| \geq n - \log(1/\delta)$ and $\tilde{\mathrm{H}}_\infty(Y|X, Z) \geq \tilde{\mathrm{H}}_\infty(X, Y|Z) - \log|\mathrm{Supp}(X)| \geq n' - \log(1/\delta)$.

If Ext_{out} is not average-case, take Z to be a constant independent of X and Y, and if Ext_{out} is average-case then take the average of both sides over Z. The claim for non-strong Ext_{in} then follows by taking $S' = U_d$ which bounds the first term by $(1 + 1/\alpha) \cdot \varepsilon_{in}$ and the second by ε_{out}. The claim for strong Ext_{in} follows by choosing $S' = U_{\{s\}}$ to be the point mass on $s \in \{0,1\}^d$ and then taking the expectation of both sides over a uniform $s \in \{0,1\}^d$. □

Remark 5.5. Theorem 5.4 in fact a construction of a *block-source* KL-extractor, meaning that the claimed error bounds hold for any joint distributions (X, Y) such that $\mathrm{H}_\infty(Y) \geq n' - \log(1/\delta)$ and $\tilde{\mathrm{H}}_\infty(X|Y) \geq n - \log(1/\delta)$ rather than just those distributions with $\mathrm{H}_\infty(X, Y) \geq n + n' - \log(1/\delta)$. The extra $\log(1/\delta)$ entropy loss inherent in the non-block analysis is why Reingold, Wigderson, and Vadhan [RVW00] introduced the zig-zag product for extractors, which we will apply for KL-extractors in Corollary 5.19.

5.2 Existing explicit constructions

The construction of Theorem 5.4 required both a $\mathrm{D}_{1+\alpha}$-extractor and an average-case KL-extractor, so for the result not to be vacuous we need to show the existence of such extractors. Thankfully, Example 2.4 implies that extractors for ℓ_2 are also extractors for D_2, so we can use existing ℓ_2 extractors from the literature, such as the Leftover Hash Lemma of Impagliazzo, Levin, and Luby [ILL89] (see also [McI87, BBR88]) and its variant using almost-universal hash functions due to Srinivasan and Zuckerman [SZ99].

Proposition 5.6 ([McI87, BBR88, ILL89, IZ89, SZ99, DORS08])**.** *Let $\mathcal{H}$ be a collection of ε-almost universal hash functions from the set $\{0,1\}^n$ to the set $\{0,1\}^m$, meaning that for all $x \neq y \in \{0,1\}^n$ it holds that $\Pr_{h \sim \mathcal{H}}[h(x) = h(y)] \leq (1 + \varepsilon)/2^m$. Then the function* $\mathrm{Ext} : \{0,1\}^n \times \mathcal{H} \to \mathcal{H} \times \{0,1\}^m$ *defined by* $\mathrm{Ext}(x, h) = (h, h(x))$ *is an average-case $(m + \log(1/\varepsilon), 2/\ln 2 \cdot \varepsilon)$ D_2-extractor.*

In particular, for every $k, n \in \mathbb{N}$ and $1 > \varepsilon > 0$ there is an explicit strong average-case (k, ε) extractor for D_2 *(and KL) with seed length $d = O(k + \log(n/\varepsilon))$ and $m = k - \log(1/\varepsilon) - O(1)$, given by* $\mathrm{Ext}'(x, h) = h(x)$ *for h drawn from an appropriate almost-universal hash family.*

Proof. The D_2 claim is implicit in Rackoff's proof of the Leftover Hash Lemma (see [IZ89]) and Srinivasan and Zuckerman's proof of the claim for total variation [SZ99], which both analyzed the *collision probability* of the output, and the average-case claim was proved by Dodis, Ostrovsky, Reyzin, and Smith [DORS08], though we include a proof here for completeness.

Given a joint distribution (Z, X) such that X is distributed over $\{0,1\}^n$ with $\tilde{\mathrm{H}}_\infty(X|Z) \geq m + \log(1/\varepsilon)$, we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{z\sim Z}[\mathrm{D}_2(\mathrm{Ext}(X|_{Z=z},\mathcal{H}) \parallel \mathcal{H}\times U_m)] \\
&\quad= \mathop{\mathbb{E}}_{z\sim Z}\left[\log\left(2^m\cdot|\mathcal{H}|\cdot \Pr_{h,h'\sim\mathcal{H},x,x'\sim X|_{Z=z}}[(h,h(x))=(h',h'(x'))]\right)\right] \\
&\quad= \mathop{\mathbb{E}}_{z\sim Z}\left[\log\left(2^m\cdot \Pr_{h\sim\mathcal{H},x,x'\sim X|_{Z=z}}\left[x=x'\vee\left(x\neq x'\wedge h(x)=h(x')\right)\right]\right)\right] \\
&\quad\le \mathop{\mathbb{E}}_{z\sim Z}\left[\log\left(2^m\cdot\left(2^{-\mathrm{H}_\infty(X|_{Z=z})}+\frac{1+\varepsilon}{2^m}\right)\right)\right] \\
&\quad\le \log\left(\mathop{\mathbb{E}}_{z\sim Z}\left[2^{m-\mathrm{H}_\infty(X|_{Z=z})}\right]+1+\varepsilon\right) && \text{(by Jensen's inequality)} \\
&\quad= \log\left(2^{m-\tilde{\mathrm{H}}_\infty(X|Z)}+1+\varepsilon\right)\le\log(1+2\varepsilon)\le\frac{2}{\ln 2}\cdot\varepsilon.
\end{aligned}
$$

The in particular statement follows from Lemma 5.7 below and from the existence of ε-almost universal hash families with size $\mathrm{poly}(2^k, n, 1/\varepsilon)$ as constructed by [SZ99]. ☐

To establish the claim about strong extractors, we generalize Lemma 5.2 to extractors for $\mathrm{D}_{1+\alpha}$ for $\alpha > 0$:

Lemma 5.7. *If* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^d \times \{0,1\}^m$ *is a* (k,ε) $\mathrm{D}_{1+\alpha}$*-extractor (respectively average-case* $\mathrm{D}_{1+\alpha}$*-extractor) for* $\alpha > 0$ *such that* $\mathrm{Ext}(x,s) = (s, \mathrm{Ext}'(x,s))$*, then* Ext' *is a strong* (k,ε) $\mathrm{D}_{1+\alpha}$*-extractor (respectively strong average-case* (k,ε) $\mathrm{D}_{1+\alpha}$*-extractor).*

Proof.

$$
\begin{aligned}
\mathop{\mathbb{E}}_{s\sim U_d}\left[\mathrm{D}_{1+\alpha}\left(\mathrm{Ext}'(X,s)\parallel U_m\right)\right] &= \mathop{\mathbb{E}}_{s\sim U_d}\left[\frac{1}{\alpha}\log\left(1+2^{m\alpha}\sum_{y\in\{0,1\}^m}\Pr\left[\mathrm{Ext}'(X,s)=y\right]^{1+\alpha}\right)\right] \\
&\le \frac{1}{\alpha}\log\left(1+2^{m\alpha}\mathop{\mathbb{E}}_{s\sim U_d}\left[\sum_{y\in\{0,1\}^m}\Pr\left[\mathrm{Ext}'(X,s)=y\right]^{1+\alpha}\right]\right) \\
&= \frac{1}{\alpha}\log\left(1+2^{\alpha(m+d)}\sum_{(s,y)\in\{0,1\}^{d+m}}\Pr\left[(U_d,\mathrm{Ext}'(X,U_d))=(s,y)\right]^{1+\alpha}\right) \\
&= \mathrm{D}_{1+\alpha}(\mathrm{Ext}(X,U_d)\parallel U_d,U_m)
\end{aligned}
$$

☐

Following Vadhan [Vad12], we also note that the extractor based on expander walks due to Goldreich and Wigderson [GW97], which has the nice property that its seed length depends only on $n-k$ the entropy deficiency of the source rather than n itself, is also an ℓ_2 extractor. Before stating the extractor formally, we introduce some notation and terminology we will need.

Definition 5.8. Let G be a D-regular graph on $\{0,1\}^n$ with adjacency matrix A_G and transition matrix $M_G = \frac{1}{D}A_G$. Then if M_G has eigenvalues $1 = \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge -1$, the *spectral expansion* of G is $\lambda = \max\{\lambda_2, -\lambda_n\}$. A function $\Gamma_G : \{0,1\}^n \times [D] \to \{0,1\}^n$ is a *neighbor function of* G if there is some labelling of the edges of G for which $\Gamma_G(v,i)$ is the vertex obtained by following the ith edge out of v in G. Γ_G is *consistently labelled* if for all $v \neq v' \in \{0,1\}^n$ and $i \in [D]$ we have $\Gamma(v,i) \neq \Gamma(v',i)$, that is, at most one incoming edge is labelled by i.

Lemma 5.9. *Let* $\Gamma : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^n$ *be the neighbor function of a graph* G *with spectral expansion* λ*. Then for every* $0 \le k \le n$*,* Γ *is a* $\left(k, \lambda\sqrt{2^{-k}-2^{-n}}\right)$ ℓ_2*-extractor and a* $\left(k, \log\left(1+\lambda^2\left(2^{n-k}-1\right)\right)\right)$ D_2*-extractor. Furthermore, if* Γ_G *is consistently labelled, then the function* $\mathrm{W}(x,s) = s$ *is such that* (Γ_G, W) *is an injection out of* $\{0,1\}^n \times \{0,1\}^d$*.*

In particular, if $\lambda^2 \leq \varepsilon \cdot 2^{k-n}$ *then* Ext *is an average-case* $(k, \sqrt{\varepsilon} \cdot 2^{-n/2})$ ℓ_2*-extractor and an average-case* $(k, \varepsilon/\ln 2)$ D_2*-extractor.*

Proof. If X is a distribution over $\{0,1\}^n$ with $\mathrm{H}_\infty(X) \geq k$, then $\log(1 + 2^n d_{\ell_2}(X, U_n)) = \mathrm{D}_2(X \parallel U_n) \leq \mathrm{D}_\infty(X \parallel U_n) \leq n - k$ so that $d_{\ell_2}(P, U_n) \leq \sqrt{2^{-k} - 2^{-n}}$. The ℓ_2-extractor result follows since the action of Γ_G reduces the ℓ_2 distance to uniform by a factor of λ, and the D_2-extractor result from the fact that $D_2(P \parallel U_n) = \log\Big(1 + 2^n d_{\ell 2}(P, U_n)^2\Big)$ for every distribution P on $\{0,1\}^n$.

For the furthermore claim, we need to show that $(x, s) \mapsto (\Gamma_G(x, s), s)$ is an injection, or equivalently that given $\Gamma_G(x, s)$ and s, one can recover x. But by definition of consistent labelling, at most one edge into $\Gamma_G(x, s)$ is labelled by s, and so taking this edge from $\Gamma_G(x, s)$ gives x, as desired. Finally, the in particular claim follows by Jensen's inequality, since log and square-root are concave, and $\mathbb{E}_{z \sim Z}\big[2^{-\mathrm{H}_\infty(X|_{Z=z})}\big] = 2^{-\tilde{\mathrm{H}}_\infty(X|Z)}$ by definition. ☐

Remark 5.10. The fact that $(s, \mathrm{Ext}(x, s))$ is an injection implies that, unlike for the extractors from hashing of Proposition 5.6, the result of prepending the seed to the output of the expander-walk extractor does *not* give a D_2 extractor. However, it will be very useful in concert with Reingold, Vadhan, and Wigderson's zig-zag product for extractors [RVW00] to avoid the entropy loss in Theorem 5.4.

Corollary 5.11 ([GW97] [Vad12, Discussion after Theorem 6.22])**.** *There is a universal constant* $C \geq 1$ *such that for every* $1 > \varepsilon > 0$*,* $\Delta > 0$*, and* $n \in \mathbb{N}$ *there is an explicit* $(n - \Delta, \varepsilon/\ln 2)$ *average-case* D_2*-extractor (respectively* $(n - \Delta, \sqrt{\varepsilon} \cdot 2^{-n/2})$ *average-case* ℓ_2*-extractor)* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^n$ *with* $d = \lceil C \cdot (\Delta + \log(1/\varepsilon)) \rceil + O(1)$ *such that the function* $(x, s) \mapsto (s, \mathrm{Ext}(x, s))$ *is an injection.*

Moreover, if there is an explicit construction of consistently labelled neighbor functions for Ramanujan graphs over $\{0,1\}^n$ *with degree* $D = O\big(2^\Delta/\varepsilon\big)$*, then one can take* $C = 1$*.*

Proof. By Lemma 5.9 it suffices to demonstrate the existence of an explicit D-regular expander graph over $\{0,1\}^n$ with a consistently labelled neighbor function Γ_G, spectral expansion $\lambda^2 \leq \varepsilon \cdot 2^{-\Delta}$, and $D = O\Big(\big(2^\Delta/\varepsilon\big)^C\Big)$. The claim about Ramanujan graphs is thus immediate since a Ramanujan graph with degree $O(2^\Delta/\varepsilon)$ has $\lambda^2 \leq 4/D \leq \varepsilon \cdot 2^{-\Delta}$.

Without the assumption of good Ramanujan graphs, we can use a power of the the explicit constant degree expander of Margulis–Gabber–Galil [Mar73, GG81] (technically this requires n even, which following Goldreich [Gol11a] we can fix when n is odd by joining two graphs on $\{0,1\}^{n-1}$ by the canonical perfect matching, and we can add self-loops to ensure the degree is a power of 2). This graph G is consistently labelled with degree $D_{MGG} = O(1)$ and constant spectral expansion $\lambda_{MGG} < 1$. Then the graph G^w on $\{0,1\}^n$ with edges representing w-length paths has spectral expansion λ_{MGG}^w and degree D_{MGG}^w, which for $w = \lceil \log_{\lambda_{MGG}}(1/2) \cdot (\Delta + \log(1/\varepsilon)) \rceil$ gives $\lambda \leq \varepsilon \cdot 2^{-\Delta}$ and degree $D = O\Big(\big(2^\Delta/\varepsilon\big)^C\Big)$ for $C \leq \log(D_{MGG}) \cdot \log_{\lambda_{MGG}}(1/2)$ as desired. ☐

We argued that the above extractors are KL-extractors using the fact they are ℓ_2 (and thus D_2) extractors, but one can also show that any total variation extractor with sufficiently small error is a KL-extractor, albeit with some loss of parameters.

Lemma 5.12. *For every* (k, ε) *extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *for total variation distance such that* $\varepsilon \leq 1/2$*,* Ext *is also a* $(k, m \cdot \varepsilon + h(\varepsilon))$*-KL-extractor, where* $h(x) = x \log(1/x) + (1 - x) \log(1/(1 - x))$ *is the binary entropy function. Furthermore, if* Ext *is strong, average-case, or both as a total variation extractor, then it has the same properties as a KL-extractor.*

In particular, if $\varepsilon' = \frac{\min(\varepsilon, 1/2)}{48(m + \log(1/\varepsilon))}$*, then every* (k, ε') *extractor (respectively strong extractor) is an average-case* (k, ε) *KL-extractor (respectively strong average-case* (k, ε) *KL-extractor).*

Proof. The main claim is an immediate corollary of Lemmas 4.9 and 5.1. The in particular statement follows since Ext being a (k, ε') extractor (respectively strong extractor) implies by Theorem 3.12 that it is a $(k, 3\varepsilon')$ average-case (respectively strong average-case) extractor, so since we have chosen ε' to make

$m \cdot 3\varepsilon' + h(3\varepsilon') \le \varepsilon$, we know Ext is an average-case (k, ε) KL-extractor (respectively strong average-case KL-extractor). $\square$

Remark 5.13. Reducing ε by a factor of $m + \log(1/\varepsilon)$ increases the seed length and entropy loss of the input extractor. For the former, this is often (but not always) tolerable since the input extractor may already depend suboptimally on $\log(n/\varepsilon)$. For the latter, we will show in Corollary 5.21 how to use the transform of Raz, Reingold, and Vadhan [RRV02] to recover $O(\log(m/\varepsilon))$ bits of lost entropy (at least this much must be lost by Radhakrishnan and Ta-Shma [RT00]) at a cost of $O(\log(n/\varepsilon))$ in the seed length.

Instantiating Lemma 5.12 with the Guruswami–Umans–Vadhan [GUV09] extractor for total variation distance, we see that the increased seed length and entropy loss are simply absorbed into the existing hidden constants:

Theorem 5.14 (KL-analogue of [GUV09, Theorem 1.5]). *For every $n \in \mathbb{N}$, $k \le n$, and $1 > \alpha, \varepsilon > 0$, there is an explicit average-case (respectively strong average-case) (k, ε) KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with $d \le \lg n + O_\alpha(\lg(k/\varepsilon))$ and $m \ge (1-\alpha)k$ (respectively $m \ge (1-\alpha)k - O_\alpha(\log(n/\varepsilon))$).*

5.3 Reducing the entropy loss of KL-extractors

In this section, we show how to avoid the entropy loss inherent in Theorem 5.4 using the zig-zag product for extractors, introduced by Reingold, Vadhan, and Wigderson [RVW00]. This product combines a technique of Raz and Reingold [RR99] to preserve entropy and the method of Wigderson and Zuckerman [WZ99] to extract entropy left over in a source after an initial extraction, and we show that these techniques extend to the setting of KL-extractors. Furthermore, these techniques (along with the Leftover Hash Lemma) are also the key to the transformation of Raz, Reingold, and Vadhan [RRV02] to convert an arbitrary extractor into one with optimal entropy loss, so we show that this transformation works for KL-extractors as well.

For all of these results, the key is the following lemma:

Lemma 5.15 (Re-extraction from leftovers). *Let*

1. $\mathrm{Ext}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m_1}$ *be a (k_1, ε_1) KL-extractor,*
2. $\mathrm{W}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^w$ *be a function such that* $(\mathrm{Ext}_1, \mathrm{W}_1) : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m_1} \times \{0,1\}^w$ *is an injective map,*
3. $\mathrm{Ext}_2 : \{0,1\}^w \times \{0,1\}^{d_2} \to \{0,1\}^{m_2}$ *be a (k_2, ε_2) average-case KL-extractor for $k_2 \le k_1 + d_1 - m_1$.*

Then $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^{d_1+d_2} \to \{0,1\}^{m_1+m_2}$ *defined by* $\mathrm{Ext}\big(x, (s,t)\big) = \big(\mathrm{Ext}_1(x,s), \mathrm{Ext}_2\big(\mathrm{W}_1(x,s), t\big)\big)$ *is a $(k_1, \varepsilon_1 + \varepsilon_2)$ KL-extractor. Furthermore, if* Ext_1 *is average-case then so is* Ext.

Remark 5.16. The pair $(\mathrm{Ext}_1, \mathrm{W}_1)$ is a special case of what Raz and Reingold [RR99] called an *extractor-condenser pair.* One can think of W_1 as preserving "leftovers" or "waste," which is then "re-extracted" or "recycled" by Ext_2. The identity function on $\{0,1\}^n \times \{0,1\}^{d_1}$ is a valid choice of W_1, but the advantage of the more general formulation is that w can be much smaller than $n + d_1$, and most known explicit constructions of extractors have seed length depending on the input length of the source.

Proof. Given any joint distribution (Z, X) such that X is distributed over $\{0,1\}^n$ and $\tilde{\mathrm{H}}_\infty(X|Z) \ge k_1$, we have for every $z \in \mathrm{Supp}(Z)$ that

$$\begin{aligned} &\mathrm{KL}(\mathrm{Ext}(X|_{Z=z}, (U_{d_1}, U_{d_2})) \parallel U_{m_1+m_2}) \\ &\quad = \mathrm{KL}\big(\mathrm{Ext}_1(X|_{Z=z}, U_{d_1}), \mathrm{Ext}_2\big(\mathrm{W}_1(X|_{Z=z}, U_{d_1}), U_{d_2}\big) \,\big\|\, U_{m_1}, U_{m_2}\big) \\ &\quad = \mathrm{KL}(\mathrm{Ext}_1(X|_{Z=z}, U_{d_1}) \parallel U_{m_1}) \\ &\qquad + \mathop{\mathbb{E}}_{o_1 \sim \mathrm{Ext}_1(X|_{Z=z}, s)}\Big[\mathrm{KL}\Big(\mathrm{Ext}_2\Big(\mathrm{W}_1(X, U_{d_1})|_{Z=z, \mathrm{Ext}_1(X, U_{d_1})=o_1}, U_{d_2}\Big) \,\Big\|\, U_{m_2}\Big)\Big] \end{aligned} \tag{5.16.1}$$

where the last line follows from the chain rule for KL divergence. Note that

$$\begin{aligned}
&\tilde{\mathrm{H}}_\infty\Big(\mathrm{W}_1\big(X,U_{d_1}\big)\ \Big|\ Z,\mathrm{Ext}_1\big(X,U_{d_1}\big)\Big)\\
&\qquad = \tilde{\mathrm{H}}_\infty\Big(\mathrm{Ext}_1\big(X,U_{d_1}\big),\mathrm{W}_1\big(X,U_{d_1}\big)\ \Big|\ Z,\mathrm{Ext}_1\big(X,U_{d_1}\big)\Big)\\
&\qquad = \tilde{\mathrm{H}}_\infty\Big(X,U_{d_1}\ \Big|\ Z,\mathrm{Ext}_1\big(X,U_{d_1}\big)\Big) && ((\mathrm{Ext}_1,\mathrm{W}_1)\text{ is an injection})\\
&\qquad \geq \tilde{\mathrm{H}}_\infty(X,U_{d_1}\mid Z) - \log\big|\mathrm{Supp}\big(\mathrm{Ext}_1\big(X,U_{d_1}\big)\big)\big| && (*)\\
&\qquad = \tilde{\mathrm{H}}_\infty(X\mid Z) + \mathrm{H}_\infty(U_{d_1}) - \log\big|\mathrm{Supp}\big(\mathrm{Ext}_1\big(X,U_{d_1}\big)\big)\big| && (\text{by independence})\\
&\qquad \geq k_1+d_1-m_1 \geq k_2
\end{aligned}$$

where the line (*) follows from standard properties of conditional min-entropy (e.g. [DORS08, Lemma 2.2]). That Ext is a $(k_1,\varepsilon_1+\varepsilon_2)$ KL-extractor now follows immediately from Eq. (5.16.1) by taking Z independent of X, and the average-case claim follows from taking expectations over $z\sim Z$. ☐

Remark 5.17. The proof above in fact works any weak divergence D such that $\mathrm{D}(X,Y\parallel U_{m_1},U_{m_2})\leq \mathrm{D}(X\parallel U_{m_1})+\mathbb{E}_{x\sim X}[\mathrm{D}(Y|_{X=x}\parallel U_{m_2})]$ for all joint distributions (X,Y) independent of (U_{m_1},U_{m_2}). In particular, the proof also gives Lemma 5.15 for standard (total variation) extractors.

By Lemma 5.2, we get an analogous result for strong KL-extractors.

Corollary 5.18. *Let*

1. $\mathrm{Ext}_1:\{0,1\}^n\times\{0,1\}^{d_1}\to\{0,1\}^{m_1}$ *be a strong* (k_1,ε_1) *KL-extractor,*
2. $\mathrm{W}_1:\{0,1\}^n\times\{0,1\}^{d_1}\to\{0,1\}^w$ *be a function such that the map* $(x,s)\mapsto(s,\mathrm{Ext}_1(x,s),\mathrm{W}_1(x,s))$ *is an injection,*
3. $\mathrm{Ext}_2:\{0,1\}^w\times\{0,1\}^{d_2}\to\{0,1\}^{m_2}$ *be a* (k_2,ε_2) *strong average-case KL-extractor for* $k_2\leq k_1-m_1$.

Then $\mathrm{Ext}:\{0,1\}^n\times\{0,1\}^{d_1+d_2}\to\{0,1\}^{m_1+m_2}$ *defined by* $\mathrm{Ext}\big(x,(s,t)\big)=\big(\mathrm{Ext}_1(x,s),\mathrm{Ext}_2\big(\mathrm{W}_1(x,s),t\big)\big)$ *is a strong* $(k_1,\varepsilon_1+\varepsilon_2)$ *KL-extractor. Furthermore, if* Ext_1 *is average-case then so is* Ext.

The zig-zag product for extractors due to Reingold, Vadhan, and Wigderson [RVW00] (in the special case of injective (Ext, W)-pairs) is an immediate consequence of Lemma 5.15 and Theorem 5.4 our basic composition result. Recall that Theorem 5.4 was able to combine an "outer" extractor, generally taken to have seed length depending only (but linearly) on $n-k$, with an "inner" extractor to produce seeds for the outer extractor with logarithmic seed length. However, as discussed in Remark 5.5 that basic composition necessarily lost $\log(1/\delta)$ bits of entropy, so the zig-zag product uses Lemma 5.15 to recover this entropy, using an (Ext, W)-pair to ensure that the re-extraction adds additional seed length depending logarithmically on $n-k$ rather than n.

Corollary 5.19 (Zig-zag product for KL-extractors, analogous to [RVW00, Theorem 3.6])**.** *Let*

1. $\mathrm{Ext}_{out}:\{0,1\}^n\times\{0,1\}^d\to\{0,1\}^m$ *be an* $(n-\log(1/\delta),\varepsilon_{out})$ *extractor for* $\mathrm{D}_{1+\alpha}$ *with* $\alpha>0$,
2. $\mathrm{W}_{out}:\{0,1\}^n\times\{0,1\}^d\to\{0,1\}^w$ *be a function such that the pair* $(\mathrm{Ext}_{out},\mathrm{W}_{out})$ *is an injection from* $\{0,1\}^n\times\{0,1\}^d$,
3. $\mathrm{Ext}_{in}:\{0,1\}^{n'}\times\{0,1\}^{d'}\to\{0,1\}^d$ *be an* $(n'-\log(1/\delta),\varepsilon_{in})$ *average-case KL-extractor,*
4. $\mathrm{W}_{in}:\{0,1\}^{n'}\times\{0,1\}^{d'}\to\{0,1\}^{w'}$ *be such that the pair* $(\mathrm{Ext}_{in},\mathrm{W}_{in})$ *is an injection from* $\{0,1\}^{n'}\times\{0,1\}^{d'}$,
5. $\mathrm{Ext}_{waste}:\{0,1\}^{w+w'}\times\{0,1\}^{d''}\to\{0,1\}^{m''}$ *be an average-case* $(n+n'-\log(1/\delta)-m,\varepsilon_{waste})$ *KL-extractor,*

and define

1. $\mathrm{Ext}_{comp} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \to \{0,1\}^m$ *by* $\mathrm{Ext}_{comp}\big((x,y),s\big) = \mathrm{Ext}_{out}\big(x, \mathrm{Ext}_{in}(y,s)\big)$ *as in Theorem 5.4,*

2. $\mathrm{W}_{comp} : \{0,1\}^{n+n'} \times \{0,1\}^{d'} \to \{0,1\}^{w+w'}$ *by* $\mathrm{W}_{comp}((x,y),s) = (\mathrm{W}_{out}(x, \mathrm{Ext}_{in}(y,s)), \mathrm{W}_{in}(y,s))$,

3. $\mathrm{Ext} : \{0,1\}^{n+n'} \times \{0,1\}^{d'+d''} \to \{0,1\}^{m+m''}$ *by*

$$\mathrm{Ext}\big((x,y),(s,t)\big) = \left(\mathrm{Ext}_{comp}\big((x,y),s\big), \mathrm{Ext}_{waste}\Big(\mathrm{W}_{comp}\big((x,y),s\big),t\Big)\right)$$

as in Lemma 5.15.

Then Ext *is an* $(n+n'-\log(1/\delta), \varepsilon_{out} + (1+1/\alpha)\cdot\varepsilon_{in} + \varepsilon_{waste})$*-extractor for KL. Furthermore, if* Ext_{in} *and* Ext_{waste} *are strong average-case KL-extractors, then* Ext *is a strong KL-extractor, and if* Ext_{out} *is average-case then so is* Ext*.*

Proof. We claim that W_{comp} is such that $(\mathrm{Ext}_{comp}, \mathrm{W}_{comp})$ is an injection: by assumption on $(\mathrm{Ext}_{out}, \mathrm{W}_{out})$ we have that given $\mathrm{Ext}_{out}(x, \mathrm{Ext}_{in}(y,s))$ and $\mathrm{W}_{out}(x, \mathrm{Ext}_{in}(y,s))$ we can recover x and $\mathrm{Ext}_{in}(y,s)$, and by assumption on $(\mathrm{Ext}_{in}, \mathrm{W}_{in})$ given $\mathrm{Ext}_{in}(y,s)$ and $\mathrm{W}_{in}(y,s)$ we can recover (y,s), so that $(\mathrm{Ext}_{comp}, \mathrm{W}_{comp})$ has an inverse and is injective as desired. Therefore, since Theorem 5.4 implies Ext_{comp} is an $(n+n'-\log(1/\delta), \varepsilon_{out} + (1+1/\alpha)\cdot\varepsilon_{in})$ KL-extractor, the result follows from Lemma 5.15. The furthermore claims follow from the corresponding claims of these lemmas (and Corollary 5.18 for the strong case). ☐

Remark 5.20. Corollary 5.19 was presented by Reingold, Vadhan, and Wigderson [RVW00] as a transformation that combined three extractor-condenser pairs into a new extractor-condenser pair. We do not use this generality, so for simplicity we do not present it here, but both Lemma 5.15 and Corollary 5.19 can be easily extended in this manner if required.

The Raz–Reingold–Vadhan [RRV02] transformation to avoid entropy loss follows similarly using the Leftover Hash Lemma (Proposition 5.6).

Corollary 5.21 (KL-extractor analogue of [RRV02, Lemma 28])**.** *Let* $\mathrm{Ext}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^{m_1}$ *be a strong* $(k, \varepsilon/2)$ *KL-extractor with entropy loss* Δ_1*, meaning* $m_1 = k - \Delta_1$*. Then for every* $d_{extra} \le \Delta_1$ *there is an explicit* (k,ε) *strong KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^{d'} \to \{0,1\}^{m'}$ *with seed length* $d' = d_1 + O(d_{extra} + \log(n/\varepsilon))$ *and entropy loss* $\Delta_1 - d_{extra} + \log(1/\varepsilon) - O(1)$*, meaning* $m' = k - (\Delta_1 - d_{extra}) - \log(1/\varepsilon) + O(1)$*, which is computable in polynomial time making one oracle call to* Ext_1*. Furthermore, if* Ext_1 *is average-case then so is* Ext*.*

In particular, by taking $d_{extra} = \Delta_1$ *we get an extractor with optimal entropy loss* $\log(1/\varepsilon) + O(1)$ *by paying an additional* $O(\Delta + \log(n/\varepsilon))$ *in seed length.*

Proof. Let $\mathrm{W}_1 : \{0,1\}^n \times \{0,1\}^{d_1} \to \{0,1\}^n$ be given by $\mathrm{W}_1(x,s) = x$, and let $\mathrm{Ext}_2 : \{0,1\}^n \times \{0,1\}^{d_2} \to \{0,1\}^{m_2}$ be the strong average-case $(d_{extra}, \varepsilon/2)$ KL-extractor of Proposition 5.6 using almost-universal hash functions, so that $d_2 = O(d_{extra} + \log(n/\varepsilon))$ and $m_2 = d_{extra} - \log(1/\varepsilon) - O(1)$. The result follows from taking Ext to be the extractor of Corollary 5.18. ☐

Remark 5.22. An analogous versions of the above claim for non-strong KL-extractors follows by taking $\mathrm{W}_1(x,s) = (x,s)$ and using Lemma 5.15.

We can apply Corollary 5.21 to Theorem 5.14 the KL-extractors from the total variation extractors of Guruswami, Umans, and Vadhan [GUV09], thereby avoiding the extra $O(\log(n/\varepsilon))$ entropy loss in the strong extractors.

Corollary 5.23. *For every* $n \in \mathbb{N}$*,* $1 > \alpha, \varepsilon > 0$*, and* $k, k' \ge 0$ *with* $k + k' \le n$*, there is an explicit strong average-case* $(k+k', \varepsilon)$ *KL-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with* $d \le O_\alpha(\log(n/\varepsilon)) + O(k')$ *and* $m \ge (1-\alpha)k + k' - \log(1/\varepsilon) - O(1)$*.*

5.4 Lower bounds

In this section, we give lower bounds on extractors for the Rényi divergences D_β of all orders, including the special case $\beta = 1$ of KL-extractors. A reader primarily interested in explicit constructions of subgaussian samplers can skip to Section 6.

For Rényi divergences D_β with $\beta \le 1$ we reduce to Radhakrishnan and Ta-Shma's [RT00] lower bounds for total variation extractors and *dispersers*, which can be understood as a one-sided relaxation of total variation extractors.

Definition 5.24 (Sipser [Sip88], Cohen and Wigderson [CW89])**.** A function $\mathrm{Disp} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a (k, ε) *disperser* if for all random variables X over $\{0,1\}^n$ with $\mathrm{H}_\infty(X) \ge k$, it holds that $|\mathrm{Supp}(\mathrm{Disp}(X, U_d))| \ge (1-\varepsilon)2^m$.

Dispersers are of interest in the context of Rényi extractors because the Rényi 0-entropy of a random variable is the logarithm of its support size (see Example 2.4), and hence dispersers are equivalent to D_0-extractors:

Lemma 5.25. Disp *is a* (k, ε) *disperser if and only if* Disp *is a* $\big(k, \log\big(1/(1-\varepsilon)\big)\big)$ D_0*-extractor.*

Given Lemma 5.25, we can use the Radhakrishnan and Ta-Shma [RT00] lower bounds to give an optimal lower bound on the seed length of D_β-extractors for $\beta \le 1$ in terms of the error ε, input length n and supported entropy k (we will give a matching non-explicit upper bound in the next section), as well as lower bounds on the entropy loss. For the case $\beta = 1$ of KL-extractors, the non-explicit upper bound (Theorem 5.30) also shows that the entropy loss lower bound is optimal.

Theorem 5.26. *Let* $0 \le \beta \le 1$ *and* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a* (k, ε) *extractor for* D_β *with* $k \le n-2$, $d \le m-1$, *and* $2^{2-m} < \varepsilon < 1/4$. *Then* $d \ge \log(n-k) + \log(1/\varepsilon) - O(1)$ *and* $m \le k + d - \log\log(1/\varepsilon) + O(1)$. *Furthermore, if* ε *is at most* $\beta/(2\ln 2)$ *then* $m \le k + d - \log(1/\varepsilon) + \log(1/\beta) + O(1)$.

Proof. Since D_β is nondecreasing in β we have that Ext is a (k, ε) extractor for D_0, and thus by Lemma 5.25 it is a $(k, 1-2^{-\varepsilon})$ disperser. Then the disperser seed length lower bound of Radhakrishnan and Ta-Shma [RT00] tells us that $d \ge \log(n-k) + \log(1/(1-2^{-\varepsilon})) - O(1) \ge \log(n-k) + \log(1/\varepsilon) - O(1)$ and $m \le k + d - \log\log(1/(1-2^{-\varepsilon})) + O(1) \le k + d - \log\log(1/\varepsilon) + O(1)$.

For the other entropy loss lower bound, we use Gilardoni's [Gil10] generalization of Pinsker's inequality, which shows in particular that $d_{TV}(P, U_m) \le \sqrt{\ln 2/(2\beta) \cdot \mathrm{D}_\beta(P \,\|\, U_m)}$. Thus, Ext is also a $(k, \sqrt{\varepsilon \cdot \ln 2/(2\beta)})$ total variation extractor, and if $\sqrt{\varepsilon \cdot \ln 2/(2\beta)} \le 1/2$ (equivalently $\varepsilon \le \beta/(2\ln 2)$) then the [RT00] total variation extractor entropy loss lower bound implies that $m \le k + d - 2\log(1/\sqrt{\varepsilon \cdot \ln 2/(2\beta)}) + O(1) \le k + d - \log(1/\varepsilon) + \log(1/\beta) + O(1)$. ☐

Remark 5.27. For the case of $0 < \beta < 1$, we do not know whether the entropy loss lower bound of Theorem 5.26 is tight.

It is well-known that ℓ_2-extractors (which are equivalent to D_2-extractors by Example 2.4) require seed length at least linear in $\min(n-k, m)$ (see e.g. [Vad12, Problem 6.4]). We generalize this to give a linear seed length lower bound on D_β extractors for all $\beta > 1$, in the regime of constant ε, improving on the logarithmic lower bound given by Theorem 5.26.

Theorem 5.28. *Let* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a* $(k, 0.99)$ $\mathrm{D}_{1+\alpha}$*-extractor for* $\alpha > 0$. *Then* $d \ge \min\big\{(n-k-3)\cdot\alpha, (m-2)\cdot\alpha/(\alpha+1)\big\}$.

Proof. We follow the strategy suggested by Vadhan [Vad12, Problem 6.4], and view Ext as a bipartite graph with $N = \{0,1\}^n$ left-vertices, $M = \{0,1\}^m$ right-vertices, and $D = 2^d$ edges per left-vertex given by $E = \{(x \in \{0,1\}^n, y \in \{0,1\}^m) \mid \exists s \in \{0,1\}^d : \mathrm{Ext}(x,s) = y\}$.

Assume for the sake of contradiction that $d \le \alpha/(\alpha+1)\cdot(m-2)$ and $d \le \alpha(n-k-3)$, so that $M \ge 4D^{1+1/\alpha}$ and $N/(8D^{1/\alpha}) \ge K$. Now, we claim there exists a set $T \subseteq \{0,1\}^m$ of size at most $M/(2D^{1+1/\alpha})$ such that $X = \{x \in \{0,1\}^n \mid \exists s \in \{0,1\}^d \text{ s.t } \mathrm{Ext}(x,s) \in T\}$ has size at least $N/(8D^{1/\alpha}) \ge K$. This follows from the following iterative procedure: until $|X| \geq N/(8D^{1/\alpha})$, choose the vertex $y \in \{0,1\}^m$ of highest degree, add it to T, and remove y and its neighbors from the graph (the neighbors go in X). Then at each step we will add to X a number of vertices at least the average degree

$$\frac{(N-|X|)\cdot D}{M-|T|} \geq \frac{(N-N/(8D^{1/\alpha}))\cdot D}{M} \geq \frac{ND}{2M},$$

so that the size of T will be at most $\lceil N/(8D^{1/\alpha})\cdot 2M/ND\rceil = \lceil M/(4D^{1+1/\alpha})\rceil \leq M/(2D^{1+1/\alpha})$ as desired. Now, since X has size at least K and Ext is a $(k, 0.99)$ $\mathrm{D}_{1+\alpha}$-extractor, we have that

$$\begin{aligned}
0.99 &\geq \mathrm{D}_{1+\alpha}(\mathrm{Ext}(U_X, U_d) \parallel U_m) \\
&= \frac{1}{\alpha}\log\left(\sum_{y\in\{0,1\}^m} \frac{\Pr[\mathrm{Ext}(U_X,U_D)=y]^{1+\alpha}}{2^{-m\alpha}}\right) \\
&\geq \frac{1}{\alpha}\log\left(M^\alpha \sum_{y\in T}\Pr[\mathrm{Ext}(U_X,U_D)=y]^{1+\alpha}\right) \\
&\geq \frac{1}{\alpha}\log\left(M^\alpha\cdot|T|^{-\alpha}\cdot\left(\sum_{y\in T}\Pr[\mathrm{Ext}(U_X,U_d)=y]\right)^{1+\alpha}\right) && \text{(By Hölder's inequality)} \\
&\geq \frac{1}{\alpha}\log\Big(M^\alpha\cdot(M/(2D^{1+1/\alpha}))^{-\alpha}\cdot(1/D)^{1+\alpha}\Big) = 1 && \text{(By definition of } T\text{)}
\end{aligned}$$

which is a contradiction, as desired. ☐

We can also use this lower bound to get a similar lower bound for $d_{\ell_{1+\alpha}}$-extractors for all $\alpha > 0$, though in this case the lower bound applies up to an error threshold that depends on α.

Corollary 5.29. *Let* $\mathrm{Ext}: \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *be a* $\big(k, \varepsilon_\alpha \cdot 2^{-m\alpha/(1+\alpha)}\big)$ *extractor for* $d_{\ell_{1+\alpha}}$ *where* $\alpha > 0$ *and* $\varepsilon_\alpha = (2/3)\cdot\alpha/(\alpha+1)$*. Then* $d \geq \min\big\{(n-k-3)\cdot\alpha, (m-2)\cdot\alpha/(\alpha+1)\big\}$*.*

Proof. Note that the proof of Theorem 5.28 gave a lower bound on the sum $\sum_{y\in\{0,1\}^m} P_y^{1+\alpha}$ where $P = \mathrm{Ext}(U_X, U_d)$, whereas $d_{\ell_{1+\alpha}}(P, U_m)^{1+\alpha} = \sum_{y\in\{0,1\}^m}|P_y - 2^{-m}|^{1+\alpha}$. For ℓ_2 these can be related without any loss, but in general we can use the triangle inequality to get

$$\mathrm{D}_{1+\alpha}(P \parallel U_m) \leq \frac{1}{\alpha}\cdot\log\Big(2^{m\alpha}\cdot\Big(d_{\ell_{1+\alpha}}(P,U_m) + 2^{-m\alpha/(\alpha+1)}\Big)^{1+\alpha}\Big)$$

so that if $d_{\ell_{1+\alpha}}(P, U_m) \leq \varepsilon_\alpha \cdot 2^{-m\alpha/(1+\alpha)}$ where $\varepsilon_\alpha = (2/3)\cdot\alpha/(\alpha+1) \leq 2^{0.99\cdot\alpha/(\alpha+1)} - 1$, then $\mathrm{D}_{1+\alpha}(P \parallel U_m) \leq 0.99$, and we conclude by Lemma 5.1 and Theorem 5.28. ☐

5.5 Non-explicit construction

In this section, we show non-constructively the existence of KL-extractors matching the lower-bound of Theorem 5.26 and in particular implying the optimal parameters of standard extractors for total variation distance. Formally, we will prove:

Theorem 5.30. *For every* $n \in \mathbb{N}$*,* $k \leq n$*, and* $1 > \varepsilon > 0$ *there is an average-case (respectively strong average-case)* (k, ε) *KL-extractor* $\mathrm{Ext}: \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with seed length* $d = \log(n-k+1) + \log(1/\varepsilon) + O(1)$ *and output length* $m = k + d - \log(1/\varepsilon) + O(1)$ *(respectively* $m = k - \log(1/\varepsilon) - O(1)$*).*

Remark 5.31. For $\varepsilon \gg 1$ the above parameters are not necessarily optimal, and it would be interested to get matching upper and lower bounds in this regime of parameters.

We will prove Theorem 5.30 using the probabilistic method, analogously to Zuckerman [Zuc97] or Radhakrishnan and Ta-Shma [RT00] for total variation extractors. However, rather than using Hoeffding's inequality, we use the following lemma:

Lemma 5.32. *Let X be uniform over a subset of $\{0,1\}^n$ of size K. Then if* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *is a random function, it holds for every $\varepsilon > 0$ that*

$$\Pr_{\mathrm{Ext}}\left[\mathop{\mathbb{E}}_{s\sim U_d}[\mathrm{KL}(\mathrm{Ext}(X,s) \parallel U_m)] > \varepsilon\right] \le 2^{MD-KD\varepsilon/3}$$

where $D = 2^d$ and $M = 2^m$.

Remark 5.33. For total variation extractors, the analogous bound is

$$\Pr_{\mathrm{Ext}}\left[d_{TV}\big((U_d, \mathrm{Ext}(X,U_d)), (U_d, U_m)\big) > \varepsilon\right] \le 2^{MD-2KD\varepsilon^2/\ln 2}.$$

One sees that the bounds are very similar, except the KL divergence version depends on ε rather than ε^2. For the regime where $\varepsilon < 1$ the linear dependence is preferable, and is responsible for the $1 \cdot \log(1/\varepsilon)$ seed length for KL-extractors compared to the $2 \cdot \log(1/\varepsilon)$ seed length for total variation extractors.

Proof of Lemma 5.32. Note that for each $s \in \{0,1\}^d$ and fixed Ext, the random variable $\mathrm{Ext}(X,s)$ is uniform over the multiset $\{\mathrm{Ext}(x,s) \mid x \in \mathrm{Supp}(X)\}$. Hence, since Ext is a random function, this multiset is distributed exactly as taking K iid uniform samples from $\{0,1\}^m$, so we wish to bound the KL divergence between this empirical distribution and the true distribution. For this, the author [Agr19] gave the moment generating function bound

$$\mathop{\mathbb{E}}_{\mathrm{Ext}}\left[2^{t\cdot \mathrm{KL}(\mathrm{Ext}(X,s) \parallel U_m)}\right] \le \left(\frac{2^{t/K}}{1-t/K}\right)^{M-1}$$

for every $0 \le t < K$, which for $t = K/3$ is at most 2^M. Then since $\mathrm{Ext}(X,s)$ is independent across $s \in \{0,1\}^d$, we have

$$\begin{aligned}\Pr_{\mathrm{Ext}}\left[\mathop{\mathbb{E}}_{s\sim U_d}[\mathrm{KL}(\mathrm{Ext}(X,s) \parallel U_m)] > \varepsilon\right] &= \Pr_{\mathrm{Ext}}\left[2^{K/3\cdot\sum_{s\in\{0,1\}^d}\mathrm{KL}(\mathrm{Ext}(X,s) \parallel U_m)} > 2^{K/3\cdot D\varepsilon}\right] \\ &\le 2^{-KD\varepsilon/3}\cdot\prod_{i=1}^{D} 2^M \qquad \square\end{aligned}$$

We can now prove Theorem 5.30:

Proof of Theorem 5.30. We will show that a random function $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ is a strong average-case (k,ε) KL-extractor with positive probability, the non-strong version then follows from Lemma 5.2. By Lemma 3.14, it is enough to prove that Ext is a strong $(k-t, 2^{t+1}/3\cdot\varepsilon)$ KL-extractor for every $t \ge 0$. To reduce the range of t we need to consider, note that it suffices to be a $(\log\lfloor 2^{k-t}\rfloor, 2^{t+1}/3\cdot\varepsilon)$ extractor for every $t \ge 0$, so that by rounding down it is enough to be a $(k-t, 2^t/3\cdot\varepsilon)$ strong KL-extractor for each $t \ge 0$ such that 2^{k-t} is an integer.

Now, consider a fixed $t \ge 0$ such that 2^{k-t} is an integer. Since the KL divergence is convex in its first argument and all distributions of min-entropy at least $k-t$ are convex combinations of "flat" distributions which are uniform over a set of size 2^{k-t} (Chor and Goldreich [CG88]), it suffices to analyze the behavior of Ext on such distributions. Then for every subset $X \subseteq \{0,1\}^n$ of size 2^{k-t}, Lemma 5.32 tells us that

$$\Pr_{\mathrm{Ext}}\left[\mathop{\mathbb{E}}_{s\sim U_d}[\mathrm{KL}(\mathrm{Ext}(U_X,s) \parallel U_m)] > 2^t/3\cdot\varepsilon\right] \le 2^{MD-2^{k-t}\cdot D\cdot(2^t/3\cdot\varepsilon)/3} = 2^{MD-KD\varepsilon/9}$$

where $M = 2^m$, $D = 2^d$, and $K = 2^k$. There are $\sum_{j=0}^{K} \binom{N}{j}$ such subsets X of $\{0,1\}^n$ for which we simultaneously need to establish that $\mathbb{E}_{s \sim U_d}[\mathrm{KL}(\mathrm{Ext}(U_X, s) \parallel U_m)] \le 2^t/3 \cdot \varepsilon$, so we have by a union bound that the probability that Ext is not a strong average-case (k, ε) KL-extractor is at most

$$2^{MD - KD\varepsilon/9} \cdot \sum_{j=0}^{K} \binom{N}{j} \le 2^{MD-KD\varepsilon/9} \cdot \left(\frac{Ne}{K}\right)^K = 2^{MD+K\log(Ne/K)-KD\varepsilon/9}.$$

Hence, as long as

$$MD < \frac{KD\varepsilon}{18} \qquad\qquad K\log\left(\frac{Ne}{K}\right) < \frac{KD\varepsilon}{18}$$
$$m \le k - \log(1/\varepsilon) - O(1) \qquad\qquad d \ge \log(n-k+1) + \log(1/\varepsilon) + O(1)$$

we know that a random function is a strong average-case (k, ε) KL-extractor with positive probability as desired. □

6 Constructions of subgaussian samplers

6.1 Subconstant ε and δ

The goal of this section is to establish the following theorem, which is our explicit construction of subgaussian samplers with sample complexity having no dependence on m, and with randomness complexity and sample complexity matching the best-known $[0,1]$-valued sampler when ε and δ are subconstant (up to the hidden polynomial in the sample complexity).

Theorem 6.1. *For all $m \in \mathbb{N}$, $1 > \varepsilon, \delta > 0$, and $\alpha > 0$ there exists an explicit (δ, ε) absolute averaging sampler (respectively strong absolute averaging sampler) for subgaussian and subexponential functions* Samp : $\{0,1\}^n \to (\{0,1\}^m)^D$ *with sample complexity $D = \mathrm{poly}(\log(1/\delta), 1/\varepsilon)$ and randomness complexity $n = m + (1+\alpha)\cdot\log(1/\delta)$ (respectively $n = m + (1+\alpha)\cdot\log(1/\delta) + 2\log(1/\varepsilon) + O(1)$).*

We will use essentially the same construction used for bounded samplers in this regime, namely applying the Reingold, Wigderson, and Vadhan [RVW00] zig-zag product for extractors to combine the expander extractor of Goldreich and Wigderson [GW97] and an extractor with logarithmic seed length. However, as described in detail in Section 4.1, even the basic composition used in this construction does not work for general subgaussian extractors, so we will instead use the zig-zag product for KL-extractors (Corollary 5.19) combining extractors for Rényi divergences, specifically the D_2-extractor from Corollary 5.11 and the KL-extractor from Corollary 5.23, to get the following high-entropy KL-extractor:

Theorem 6.2. *For all integers m and $1 > \alpha, \delta, \varepsilon > 0$ there is an explicit average-case (respectively strong average-case) (k, ε) KL-extractor* Ext : $\{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with $n = m + (1+\alpha)\log(1/\delta) - O(1)$ (respectively $n = m + (1+\alpha)\cdot\log(1/\delta) + \log(1/\varepsilon) + O(1)$), $k = n - \log(1/\delta)$, and $d = O_\alpha(\log(\log(1/\delta)/\varepsilon))$.*

Proof. We prove the claim for strong extractors, for the non-strong claim one can simply define $\mathrm{Ext}(x, (s,t)) = \mathrm{Ext}_{strong}((x,t), s)$ where t has length $\log(1/\varepsilon) + O(1)$.

By Corollary 5.11, there is a universal constant $C > 0$ such that for $d_{out} = \lceil C\log(1/(\delta\varepsilon))\rceil \le C\log(1/\delta) + C\log(1/\varepsilon) + 1$ there is an explicit average-case $(n_{out} - \log(1/\delta), \varepsilon/4)$ D_2-extractor $\mathrm{Ext}_{out} : \{0,1\}^{n_{out}} \times \{0,1\}^{d_{out}} \to \{0,1\}^{n_{out}}$ with $n_{out} = m - d_{out}$. Furthermore, Ext_{out} has the property that the function $\mathrm{W}_{out}(x,s) = s$ is such that $(\mathrm{Ext}_{out}, \mathrm{W}_{out})$ is an injection.

Let $k'_{in} = C\log(1/\delta)/(1-\beta)$, $k''_{in} = (C+1)\log(1/\varepsilon) + O(1)$, and $k_{in} = k'_{in} + k''_{in}$ for $0 < \beta < 1$ some parameter to be chosen later. Then by Corollary 5.23, there is an explicit $(k_{in}, \varepsilon/4)$ strong average-case KL-extractor $\mathrm{Ext}_{in} : \{0,1\}^{n_{in}} \times \{0,1\}^{d_{in}} \to \{0,1\}^{m_{in}}$ with $n_{in} = k_{in} + \log(1/\delta)$, $d_{in} = O_\beta(\log(n_{in}/\varepsilon)) + O(k''_{in}) = O_\beta(\log(\log(1/\delta)/\varepsilon))$, and $m_{in} = (1-\beta)k'_{in} + k''_{in} - \log(1/\varepsilon) - O(1) = d_{out}$. Furthermore, the function $\mathrm{W}_{in}(x,s) = (x,s)$ is an injection.

Furthermore, for $k_{waste} = (n_{out} + n_{in} - \log(1/\delta)) - n_{out} = n_{in} - \log(1/\delta) = k_{in} = k'_{in} + k''_{in}$, by Corollary 5.23 there is also an explicit $(k_{waste}, \varepsilon/4)$ strong average-case KL-extractor $\mathrm{Ext}_{waste} : \{0,1\}^{d_{out}+n_{in}+d_{in}} \times \{0,1\}^{d_{waste}} \to \{0,1\}^{m_{waste}}$ such that $m_{waste} = d_{out}$ and $d_{waste} = O_\beta(\log((d_{out} + n_{in} + d_{in})/\varepsilon)) + O(k''_{in}) = O_\beta(\log(\log(1/\delta)/\varepsilon))$.

Then by the zig-zag product for KL-extractors (Corollary 5.19), there is an explicit $(n_{out}+n_{in}-\log(1/\delta), \varepsilon)$ strong average-case KL-extractor $\mathrm{Ext} : \{0,1\}^{n_{out}+n_{in}} \times \{0,1\}^{d_{in}+d_{waste}} \to \{0,1\}^{n_{out}+m_{waste}}$, where we have

$$\begin{aligned} n_{out} + n_{in} &= (m - d_{out}) + \Big(\big(C\log(1/\delta)/(1-\beta) + (C+1)\log(1/\varepsilon) + O(1)\big) + \log(1/\delta)\Big) \\ &\leq m + \log(1/\delta) + \log(1/\varepsilon) + \log(1/\delta)\cdot C \cdot \big(1/(1-\beta) - 1\big) + O(1) \\ d_{in} + d_{waste} &= O_\beta(\log(\log(1/\delta)/\varepsilon)) \\ n_{out} + m_{waste} &= (m - d_{out}) + d_{out} = m. \end{aligned}$$

Choosing $\beta = \alpha/(\alpha + C)$ so that $C \cdot \big(1/(1-\beta) - 1\big) \leq \alpha$ gives the claim. □

We can now prove Theorem 6.1.

Proof of Theorem 6.1. Let $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ be the explicit $\big(n - \log\big(1/(\delta/2)\big), \varepsilon^2\big)$ KL-extractor (respectively strong KL-extractor) of Theorem 6.2, so that $d = O_\alpha(\log\log(1/\delta)/\varepsilon)$ and $n = m + (1+\alpha)\log(1/\delta)$ (respectively $n = m + (1+\alpha)\log(1/\delta) + 2\log(1/\varepsilon) + O(1)$).

Then by Lemmas 4.9 and 5.1, Ext is also an $\big(n - \log\big(1/(\delta/2)\big), \varepsilon\big)$ $d_{\mathcal{E}}$-extractor (respectively strong $d_{\mathcal{E}}$-extractor), so by Theorem 3.8 the function $\mathrm{Samp} : \{0,1\}^n \times (\{0,1\}^m)^D$ given by $\mathrm{Samp}(x)_i = \mathrm{Ext}(x, i)$ is an explicit $(\delta/2, \varepsilon)$ sampler for $\mathcal{E}$ (respectively strong sampler for $\mathcal{E}$), and thus by symmetry of $\mathcal{E}$ an explicit (δ, ε) absolute subexponential sampler (respectively absolute strong subexponential sampler) as desired. □

6.2 Constant δ

We recall from the introduction that the pairwise independent sampler of Chor and Goldreich [CG89] works for subgaussian functions, and in fact the more general class of functions with bounded variance. The sampler has exponentially worse dependence on δ than is necessary for subgaussian samplers, but is very simple and has randomness complexity optimal up to constant factors.

Theorem 6.3 ([CG89])**.** *For all $m \in \mathbb{N}$ and $1 > \varepsilon, \delta > 0$ with $1/(\delta\varepsilon^2) < 2^m$, there is an explicit strong sampler* $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ *for functions with bounded variance $\mathcal{M}_2$, with randomness complexity $n = O(m)$ and sample complexity $D = O\big(\frac{1}{\varepsilon^2\delta}\big)$ defined as* $\mathrm{Samp}(h)_d = h(d)$ *where h is drawn at random from a size 2^n pairwise-independent hash family $\mathcal{H}$ of functions from $[D] \to \{0,1\}^m$.*

Proof. The fact that pairwise independence gives rise to a strong bounded-variance sampler is immediate by Chebyshev's inequality. The existence of a pairwise indepenent hash family with the claimed parameters is due to Chor and Goldreich [CG89], with similar constructions in the probability literature due to Joffe [Jof71]. □

We also show that the Expander Neighbor sampler of [KPS85, GW97] is a bounded-variance sampler.

Theorem 6.4. *There is a universal constant $C \geq 1$ such that for all $m \in \mathbb{N}$ and $1 > \varepsilon, \delta > 0$ there is an explicit sampler* $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ *for functions with bounded variance $\mathcal{M}_2$, with randomness complexity $n = m$ and sample complexity $D = O\Big(\big(\frac{1}{\varepsilon^2\delta}\big)^C\Big)$. Moreover, if the algorithm is given access to a consistently labelled neighbor function of a Ramanujan graph over $\{0,1\}^n$ of degree $O(1/(\delta\varepsilon^2))$, then one can take $C = 1$.*

Proof. By Corollary 5.11, there is an explicit $(n - \log(1/\delta), \varepsilon \cdot 2^{-m/2})$ ℓ_2-extractor $\mathrm{Ext} : \{0,1\}^m \times \{0,1\}^d \to \{0,1\}^m$ with $d = \lceil C(\log(1/\delta) + 2\log(1/\varepsilon))\rceil + O(1)$, where one can take $C = 1$ given the assumed Ramanujan graph. Then by Proposition 2.11 Ext is also an $(n - \log(1/\delta), \varepsilon)$ $\mathcal{M}_2$-extractor, so we conclude by Theorem 3.8. □

Remark 6.5. Note that given explicit constructions of Ramanujan graphs, Theorem 6.4 has the same sample complexity but better randomness complexity than the sampler of Theorem 6.3.

6.3 Non-explicit construction

Applying Lemmas 4.9 and 5.1 to Theorem 5.30 our non-explicit construction of KL-extractors gives:

Corollary 6.6. *For every $n \in \mathbb{N}$, $k \le n$, and $1 > \varepsilon > 0$ there is an average-case (respectively strong average-case) (k, ε) $d_{\mathcal{E}}$-extractor* $\mathrm{Ext} : \{0,1\}^n \times \{0,1\}^d \to \{0,1\}^m$ *with $d = \log(n-k+1) + 2\log(1/\varepsilon) + O(1)$ and $m \ge k + d - 2\log(1/\varepsilon) - O(1)$ (respectively $m \ge k - 2\log(1/\varepsilon) - O(1)$)*

Since $d_{\mathcal{E}}$-extractors are also total variation extractors, Corollary 6.6 is optimal up to additive constants by the lower bound of Radhakrishnan and Ta-Shma [RT00].

Using the fact that extractors are samplers (Theorem 3.8), we get

Corollary 6.7. *For every integer m and $1 > \delta, \varepsilon > 0$ there is a (δ, ε) sampler (respectively strong sampler)* $\mathrm{Samp} : \{0,1\}^n \to (\{0,1\}^m)^D$ *for subgaussian and subexponential functions with sample complexity $D = O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$ and randomness complexity $n = m + \log(1/\delta) - \log\log(1/\delta) + O(1)$ (respectively $n = m + \log(1/\delta) + 2\log(1/\varepsilon) + O(1)$).*

Note that this matches the best-known (non-explicit) parameters of averaging samplers for $[0,1]$-valued functions due to Zuckerman [Zuc97].

7 Acknowledgements

The author would like to thank Jarosław Błasiok for suggesting the problem of constructing subgaussian samplers and for helpful discussions and feedback, Salil Vadhan for many helpful discussions and his detailed feedback on this writeup, and the anonymous reviewers for their helpful comments and feedback.

References

[AE05] K. M. R. Audenaert and J. Eisert, "Continuity bounds on the quantum relative entropy," *Journal of Mathematical Physics*, vol. 46, no. 10, p. 102104, Oct. 2005.

[Agr19] R. Agrawal, "Concentration of the multinomial in Kullback–Leibler divergence near the ratio of alphabet and sample sizes," *arXiv:1904.02291 [cs, math, stat]*, Apr. 2019.

[AS66] S. M. Ali and S. D. Silvey, "A General Class of Coefficients of Divergence of One Distribution from Another," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[BBR88] C. H. Bennett, G. Brassard, and J.-M. Robert, "Privacy Amplification by Public Discussion," *SIAM Journal on Computing*, vol. 17, no. 2, pp. 210–229, Apr. 1988.

[BDK+11] B. Barak, Y. Dodis, H. Krawczyk, O. Pereira, K. Pietrzak, F.-X. Standaert, and Y. Yu, "Leftover Hash Lemma, Revisited," in *Advances in Cryptology – CRYPTO 2011*, ser. Lecture Notes in Computer Science, P. Rogaway, Ed. Springer Berlin Heidelberg, 2011, pp. 1–20.

[BGG93] M. Bellare, O. Goldreich, and S. Goldwasser, "Randomness in interactive proofs," *computational complexity*, vol. 3, no. 4, pp. 319–354, Dec. 1993.

[Bła18a] J. Błasiok, Private Communication, Cambridge, MA USA, 2018.

[Bła18b] ——, "Optimal streaming and tracking distinct elements with high probability," in *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. Proceedings. Society for Industrial and Applied Mathematics, Jan. 2018, pp. 2432–2448.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, 1st ed. Oxford University Press, Feb. 2013.

[BR94] M. Bellare and J. Rompel, "Randomness-efficient oblivious sampling," in *Proceedings 35th Annual Symposium on Foundations of Computer Science*, Nov. 1994, pp. 276–287.

[CEG95] R. Canetti, G. Even, and O. Goldreich, "Lower bounds for sampling algorithms for estimating the average," *Information Processing Letters*, vol. 53, no. 1, pp. 17–25, Jan. 1995.

[CG88] B. Chor and O. Goldreich, "Unbiased Bits from Sources of Weak Randomness and Probabilistic Communication Complexity," *SIAM Journal on Computing*, vol. 17, no. 2, pp. 230–261, Apr. 1988.

[CG89] ——, "On the power of two-point based sampling," *Journal of Complexity*, vol. 5, no. 1, pp. 96–106, Mar. 1989.

[Csi63] I. Csiszár, "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten," *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, vol. 8, pp. 85–108, 1963.

[CW89] A. Cohen and A. Wigderson, "Dispersers, deterministic amplification, and weak random sources," in *30th Annual Symposium on Foundations of Computer Science*, Oct. 1989, pp. 14–19.

[DORS08] Y. Dodis, R. Ostrovsky, L. Reyzin, and A. Smith, "Fuzzy Extractors: How to Generate Strong Keys from Biometrics and Other Noisy Data," *SIAM Journal on Computing*, vol. 38, no. 1, pp. 97–139, Jan. 2008.

[DV76] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time—III," *Communications on Pure and Applied Mathematics*, vol. 29, no. 4, pp. 389–461, 1976.

[GG81] O. Gabber and Z. Galil, "Explicit constructions of linear-sized superconcentrators," *Journal of Computer and System Sciences*, vol. 22, no. 3, pp. 407–420, Jun. 1981.

[Gil98] D. Gillman, "A Chernoff Bound for Random Walks on Expander Graphs," *SIAM Journal on Computing*, vol. 27, no. 4, pp. 1203–1220, Aug. 1998.

[Gil10] G. L. Gilardoni, "On Pinsker's and Vajda's Type Inequalities for Csiszár's f-Divergences," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5377–5386, Nov. 2010.

[Gol11a] O. Goldreich, "Basic Facts about Expander Graphs," in *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, O. Goldreich, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, vol. 6650, pp. 451–464.

[Gol11b] ——, "A Sample of Samplers: A Computational Perspective on Sampling," in *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation: In Collaboration with Lidor Avigad, Mihir Bellare, Zvika Brakerski, Shafi Goldwasser, Shai Halevi, Tali Kaufman, Leonid Levin, Noam Nisan, Dana Ron, Madhu Sudan, Luca Trevisan, Salil Vadhan, Avi Wigderson, David Zuckerman*, ser. Lecture Notes in Computer Science, O. Goldreich, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 302–332.

[GUV09] V. Guruswami, C. Umans, and S. Vadhan, "Unbalanced Expanders and Randomness Extractors from Parvaresh–Vardy Codes," *Journal of the ACM*, vol. 56, no. 4, pp. 20:1–20:34, Jul. 2009.

[GV99] O. Goldreich and S. Vadhan, "Comparing Entropies in Statistical Zero Knowledge with Applications to the Structure of SZK," in *Proceedings of the Fourteenth Annual IEEE Conference on Computational Complexity*, May 1999, pp. 54–73.

[GW97] O. Goldreich and A. Wigderson, "Tiny families of functions with random properties: A quality-size trade-off for hashing," *Random Structures & Algorithms*, vol. 11, no. 4, pp. 315–343, 1997.

[ILL89] R. Impagliazzo, L. A. Levin, and M. Luby, "Pseudo-random Generation from One-way Functions," in *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, ser. STOC '89. New York, NY, USA: ACM, 1989, pp. 12–24.

[IZ89] R. Impagliazzo and D. Zuckerman, "How to recycle random bits," in *30th Annual Symposium on Foundations of Computer Science*, Oct. 1989, pp. 248–253.

[Jof71] A. Joffe, "On a sequence of almost deterministic pairwise independent random variables," *Proceedings of the American Mathematical Society*, vol. 29, pp. 381–382, 1971.

[KPS85] R. Karp, N. Pippenger, and M. Sipser, "A Time-Randomness Tradeoff," in *AMS Conference on Probabilistic Computational Complexity*, Durham, New Hampshire, 1985.

[Mar73] G. A. Margulis, "Explicit constructions of expanders," *Akademiya Nauk SSSR. Institut Problem Peredachi Informatsii Akademii Nauk SSSR. Problemy Peredachi Informatsii*, vol. 9, no. 4, pp. 71–80, 1973.

[McI87] J. L. McInnes, "Cryptography Using Weak Sources of Randomness," University of Toronto, Technical Report 194/87, 1987.

[Mor63] T. Morimoto, "Markov Processes and the H-Theorem," *Journal of the Physical Society of Japan*, vol. 18, no. 3, pp. 328–331, Mar. 1963.

[Mül97] A. Müller, "Integral Probability Metrics and Their Generating Classes of Functions," *Advances in Applied Probability*, vol. 29, no. 2, pp. 429–443, 1997.

[NZ96] N. Nisan and D. Zuckerman, "Randomness is Linear in Space," *Journal of Computer and System Sciences*, vol. 52, no. 1, pp. 43–52, Feb. 1996.

[Rén61] A. Rényi, "On Measures of Entropy and Information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[RR99] R. Raz and O. Reingold, "On recycling the randomness of states in space bounded computation," in *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing - STOC '99*. Atlanta, Georgia, United States: ACM Press, 1999, pp. 159–168.

[RRV02] R. Raz, O. Reingold, and S. Vadhan, "Extracting all the Randomness and Reducing the Error in Trevisan's Extractors," *Journal of Computer and System Sciences*, vol. 65, no. 1, pp. 97–128, Aug. 2002.

[RT00] J. Radhakrishnan and A. Ta-Shma, "Bounds for Dispersers, Extractors, and Depth-Two Superconcentrators," *SIAM Journal on Discrete Mathematics*, vol. 13, no. 1, pp. 2–24, Jan. 2000.

[RTTV08] O. Reingold, L. Trevisan, M. Tulsiani, and S. Vadhan, "New Proofs of the Green-Tao-Ziegler Dense Model Theorem: An Exposition," *arXiv:0806.0381 [math]*, Jun. 2008.

[RVW00] O. Reingold, S. Vadhan, and A. Wigderson, "Entropy waves, the zig-zag graph product, and new constant-degree expanders and extractors," in *Proceedings 41st Annual Symposium on Foundations of Computer Science*, Nov. 2000, pp. 3–13.

[Sha11] O. Shayevitz, "On Rényi measures and hypothesis testing," in *2011 IEEE International Symposium on Information Theory Proceedings*, Jul. 2011, pp. 894–898.

[Sip88] M. Sipser, "Expanders, randomness, or time versus space," *Journal of Computer and System Sciences*, vol. 36, no. 3, pp. 379–383, Jun. 1988.

[SZ99] A. Srinivasan and D. Zuckerman, "Computing with Very Weak Random Sources," *SIAM Journal on Computing*, vol. 28, no. 4, pp. 1433–1459, Jan. 1999.

[TZS06] A. Ta-Shma, D. Zuckerman, and S. Safra, "Extractors from Reed–Muller codes," *Journal of Computer and System Sciences*, vol. 72, no. 5, pp. 786–812, Aug. 2006.

[Vad12] S. P. Vadhan, *Pseudorandomness.* Boston, Mass.: Now Publishers Inc, Oct. 2012.

[vE10] T. van Erven, "When data compression and statistics disagree: Two frequentist challenges for the minimum description length principle," Ph.D. dissertation, Leiden University, 2010, oCLC: 673140651.

[vEH14] T. van Erven and P. Harremoës, "Rényi Divergence and Kullback-Leibler Divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.

[Ver18] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2018, no. 47.

[WZ99] A. Wigderson and D. Zuckerman, "Expanders That Beat the Eigenvalue Bound: Explicit Construction and Applications," *Combinatorica*, vol. 19, no. 1, pp. 125–138, Jan. 1999.

[Zol84] V. M. Zolotarev, "Probability Metrics," *Theory of Probability & Its Applications*, vol. 28, no. 2, pp. 278–302, Jan. 1984.

[Zuc97] D. Zuckerman, "Randomness-optimal oblivious sampling," *Random Structures & Algorithms*, vol. 11, no. 4, pp. 345–367, 1997.

[Zuc07] ——, "Linear Degree Extractors and the Inapproximability of Max Clique and Chromatic Number," *Theory of Computing*, vol. 3, no. 1, pp. 103–128, Aug. 2007.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il